Optimierung der Zuordnung mehrdeutiger NOESY-NMR-Signale unter Anwendung einer Datenbank nichtredundanter Proteinstrukturen

Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften (Dr. rer. Nat.)
der naturwissenschaftlichen Fakultät III – Biologie und vorklinische Medizin
der Universität Regensburg

vorgelegt von

Adel Nasser
aus Selb

im Dezember 2006

Promotionsgesuch eingereicht am: 13.12.2006
Die Arbeit wurde angeleitet von: Prof. Dr. Dr. H.R. Kalbitzer

Prüfungsausschuss:

Vorsitzender: Prof. Dr. G. Hauska
1. Gutachter: Prof. Dr. Dr. H.R. Kalbitzer
2. Gutachter: PD Dr. R. Merkl
3. Prüfer: Prof. Dr. Sterner

Zusammenfassung

In der vorliegenden Arbeit wurde die Entwicklung des Softwarepakets *AUREMOL* fortgesetzt. Zentrales Ziel des Programms ist die automatische Strukturbestimmung unter Verwendung von möglichst wenigen experimentellen Daten. Kernziel der Arbeit war die Optimierung der automatischen Zuordnung von NOESY-NMR-Spektren durch Anwendung atomspezifischer Abstandsinformation, welche aus einer großen Datenbank nichtredundanter Proteinstrukturen gewonnen wurde.

Unter Anwendung der NMR-Spektroskopie zur Strukturbestimmung, wird der Großteil der Information über interatomare Abstände im Protein aus NOESY-NMR-Spektren gewonnen. Sie stellen somit eine Schlüsselrolle bei der Strukturbestimmung dar. Ihre Auswertung gehört allerdings zu den zeitaufwendigsten und zugleich fehleranfälligsten Arbeitsschritten. Die Auswertung bzw. die Zuordnung der Spektren kann im Softwarepaket *AUREMOL* über das Programm *KNOWNOE* automatisch durchgeführt werden. Das Kernziel der Arbeit war in erster Linie die Zuordnung mehrdeutiger NOESY-Signale, die das Hauptproblem bei der Auswertung darstellen, zu verbessern. Mehrdeutige NOESY-Signale sind Signale, die aufgrund bekannter chemischer Verschiebungen nicht eindeutig einem bestimmten Atompaar zugeordnet werden konnten. Das im Programm *KNOWNOE* angewandte statistische Verfahren ist in der Lage, im Falle von zwei- oder drei unterschiedlichen Atompaaren als Zuordnungsmöglichkeit, die wahrscheinlichste Möglichkeit zu berechnen. Hierbei greift das Programm auf eine Datenbank aus Wahrscheinlichkeitsverteilungen zurück, die auf der Kenntnis interatomarer Atomabstände innerhalb bekannten Proteinstrukturen beruhen.

Ziel war es, durch Ersatz der früheren Datenbank aus Verteilungen durch eine qualitativ hochwertigere und wesentlich umfangreichere Datenbank, die Anzahl so wie die Sicherheit der Zuordnungen zu erhöhen. Die neue Datenbank wurde im Gegensatz zur früheren Datenbank aus einem Satz strukturell nichtredundanter Strukturen erzeugt. Zusätzlich unterscheidet sie sich von der früheren Datenbank durch die größere Anzahl zu Verfügung stehender Verteilungen (über 200 000 anstatt 1577) für Abstände unterschiedlicher Atompaare, durch die höhere Datenauflösung (10 000 Datenpunkte anstatt 100) der einzelnen Verteilungen, durch die größere verwendete Strukturdatenbasis (1107 Strukturen anstatt 326) bei der Generierung und durch ein akkurateres mathematische Verfahren zur Berechnung einer bestimmten Verteilungskurve (Kurvenglättung über Summierung von Gaußkurven).

Anhand der simulierten 2D-NOESY-NMR Spektren der Proteine *Tm*CSP und HPr konnte gezeigt werden, dass unter Anwendung der neuen Verteilungen die Gesamtanzahl der

erstellten Zuordnungen so wie deren Sicherheit bzw. Richtigkeit stark erhöht werden konnte. Insgesamt konnten bei vergleichbaren Versuchsbedingungen die Anteile der jeweils vorhandenen zwei- und dreideutigen NOESY-Signale, denen jeweils ein bestimmtes Atompaar mit einer hohen Wahrscheinlichkeit (z.B. 80-99%) zugewiesen wurde, in etwa verdoppelt werden. So konnten, unter Einsatz der neuen Datenbank, anstatt wie vorher etwa 20-25%, nun 40-55% der jeweils vorhandenen zwei -und dreideutigen NOESY-Signale ein bestimmtes Atompaar mit beispielsweise einer Wahrscheinlichkeit von mindestens 98% zugewiesen werden. Zugleich konnte der Anteil falsch zugewiesener Zuordnungen in der Regel um die Hälfte verringert werden. Dies ist besonders wichtig, da falsche Zuordnungen zu falschen Abstandsberechnungen führen, und dadurch zu einer Verzerrung des Strukturmodells bezüglich der wirklichen Konformation der fraglichen Proteinstruktur während der Strukturrechnung führen können. Die erreichte Minimierung falscher Zuordnungen zeigte sich besonders bei kleinen eingestellten Suchradien (<1.0 nm) im Programm *KNOWNOE* deutlich, da hierbei im Allgemeinen besonders viele falsche Zuordnungen auftreten. Der Suchradius ist ein Parameter im Programm *KNOWNOE*, der iterativ reduziert wird. Er gibt den maximalen Abstand an, den ein bestimmtes Atompaar innerhalb der bereits vorhandenen Modellstruktur haben darf, um als Zuordnungsmöglichkeit für ein bestimmtes NOESY-Signal in Frage zu kommen. So konnte der Anteil falsch zugeordneter zwei- und dreideutiger NOESY-Signale, unter dem relativ kleinen eingestellten Suchradius von beispielsweise 0,6 nm und einer eingestellten Wahrscheinlichkeitsgrenze von P=0,98 beim simulierten 2D-NOESY-Spektrum vom Protein CSP von 28,4 % auf 16,3% und beim Protein HPr von 24,5% auf 10,5% reduziert werden. Weiter konnte gezeigt werden, dass sich das hier angewandte statistische Zuordnungsverfahren bei Benutzung der neuen Verteilungen gegenüber Abstandsfehlern wesentlich toleranter verhält. So führten künstlich erzeugte Abstandfehler von beispielweise 30 % bei Anwendung der früheren Verteilungen zu erheblichen Schwankungen so wie Steigerungen der Anteile falscher Zuordnungen. Bei Benutzung der neuen Verteilungen blieben die Fehlerquoten hingegen weitgehend konstant. Dieses Ergebnis ist besonderes wichtig, da man in experimentellen NOESY-NMR-Spektren generell mit größeren Fehlern bei der Abstandbestimmung aus NOESY-NMR-Signalen rechnen muss.

Es hat sich gezeigt, dass die ereichte Steigerung der Zuordnungsanzahl vor allem auf der großen Anzahl von erzeugten Verteilungen (über 200 000) beruht. Ein weiterer wichtiger Faktor ist die stark erhöhte Datenauflösung von 10000 Datenpunkten. Die Verbesserung der Zuordnungssicherheit konnte hingegen im wesentlichem auf die größere benutzte

Strukturdatenbasis, der geringen sequentiellen Ähnlichkeit (<25%) der benutzten Proteine und dem angewandten Kurvenglättungsverfahren zurückgeführt werden.

Mit den erzeugten Datenbanken wurde in weiterem eine wertvolle Quelle struktureller Information bezüglich interatomarer Abstände zu Verfügung gestellt. Neben der Zuordnung von NOESY-NMR-Signalen, ist ihre Anwendung auch bei anderen wichtigen Arbeitschritten bei der Strukturbestimmung wie z.B. der Strukturrechnung oder der Strukturvalidierung denkbar.

Inhaltsverzeichnis

1.Einleitung

1.1 Bedeutung und Funktion von Proteinen

Proteine sind essentielle Bausteine des Lebens. Sie erfüllen im lebenden Organismus unterschiedlichste Funktionen wie z.B. Strukturgebung, Stofftransport, Katalyse biochemischer Reaktionen so wie Kommunikation zwischen Nervenzellen und Immunabwehr [1][2]. Proteine bestehen aus Ketten von Aminosäuren (Polypeptide), von denen man 20 unterschiedliche Grundtypen innerhalb der bisher bekannten Lebewesen unterscheiden kann (natürliche Aminosäuren). Typische Proteine haben ein Molekulargewicht von mehr als 10 kDa, bestehend aus jeweils einigen hundert Resten. Es gibt auch sehr kleine Proteine mit weniger als 100 Resten, wie z.B. Insulin [3], bestehend aus jeweils 50 Aminosäuren, wie auch extrem große Proteine mit mehreren Tausend Resten, wie z.B. die Glutamatdehydrogenase vom Rind [3] mit jeweils über 8300 Aminosäuren. Innerhalb lebender Organismen liegen Proteine zu etwa 70% nicht in linearer, sondern in gefalteter Form vor, und besitzen dadurch eine dreidimensionale Struktur. Die räumliche Struktur eines Proteins wird vornehmlich durch die Reihenfolge seiner Aminosäuren (Primärstruktur) bestimmt. Diese wiederum ist durch das entsprechende Gen determiniert bzw. codiert.
Proteinstrukturen besitzen einen hierarchischen Aufbau [4], wobei man generell zwischen vier strukturellen Ebenen unterscheidet (Abb. 1.2 A-D).

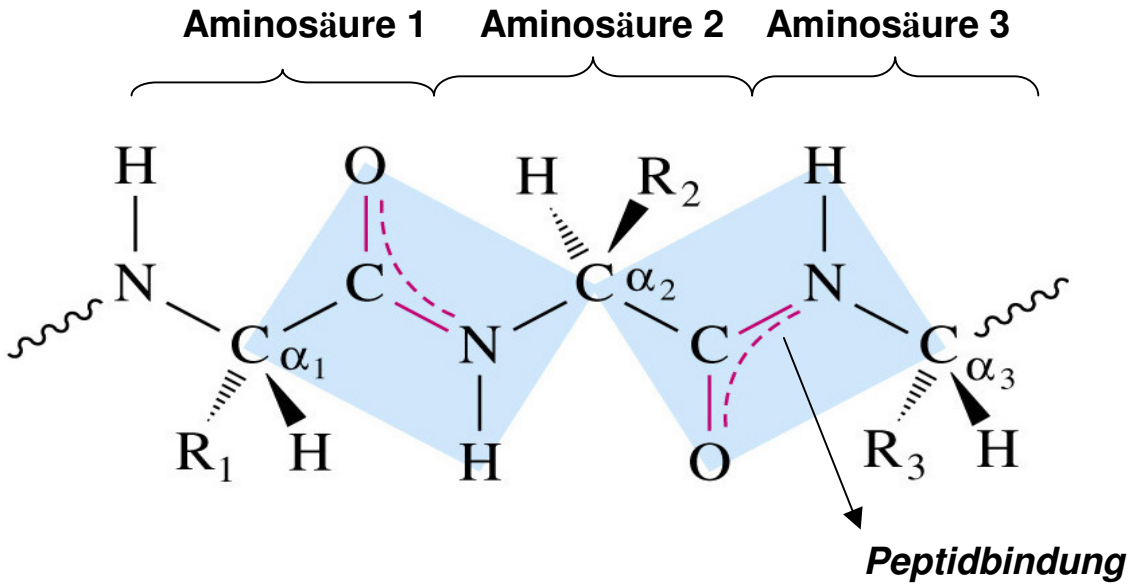


Abbildung 1.1: Verknüpfung von Aminosäuren über Peptidbindungen [6]

Die unterste Ebene stellt die Primärstruktur dar, und bezeichnet die Reihenfolge der Aminosäuren innerhalb des Proteins (Abb. 1.2 A). Die einzelnen Aminosäuren sind über sog. Peptidbindungen kovalent miteinander verknüpft (Abb. 1.1). Bei der Peptidbindung handelt es sich um eine starre planare Struktur, welche aufgrund Resonanzstabilisierung einen 40 % igen Doppelbindungscharakter besitzt [5].

Auf der zweiten Stufe der Hierarchie stehen die Sekundärstrukturelemente, die sich durch Wasserstoffbrückenbindungen zwischen den Amidgruppen und Carbonylgruppen der Aminosäuren einer Peptidkette ausbilden. Sekundärstrukturen bilden in Abhängigkeit von der Anordnung der ausgebildeten Wasserstoffbrückenbindungen entweder helikale (z.B. α-Helices) oder Faltblattstrukturen (β-Faltblätter) aus (Abb. 1.2 B).

Die nächst höhere Strukturebene wird auch als Tertiärstruktur bezeichnet (Abb. 1.2 C). Sie beschreibt die globale räumliche Faltung einer Polypeptidkette und wird vor allem durch Kontakte zwischen sequentiell weiter entfernten Aminosäuren (mehr als 5 Reste) mittels hydrophoben Wechselwirkungen oder Disulfidbrücken bestimmt.

Ist ein Protein aus mehreren Polypeptidketten oder Domänen zusammengesetzt, wird die räumliche Beziehung bzw. Anordnung der einzelnen Elemente als Quartärstruktur bezeichnet (Abb. 1.2 D).

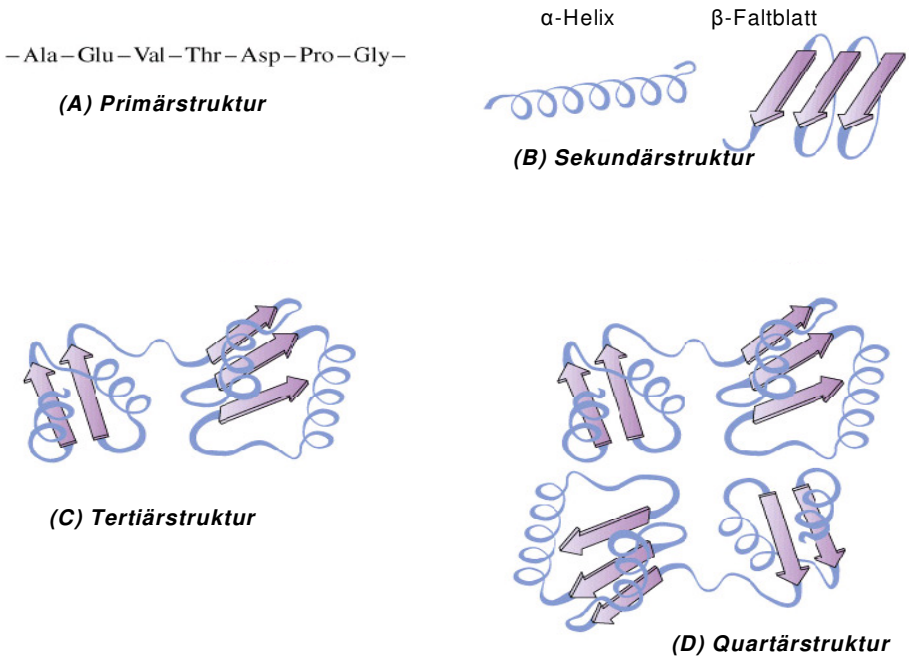


Abbildung 1.2: Hierarchischer Aufbau von Proteinen [6]

Unter physiologischen Bedingungen wird die dreidimensionale Struktur eines Proteins hauptsächlich durch elektrostatische Wechselwirkungen, Wasserstoffbrückenbindungen und hydrophobe Kräfte stabilisiert [5]. Eine Veränderung oder gar Zerstörung (Denaturierung) der Struktur eines Proteins, hervorgerufen beispielsweise durch Gendefekte, Hitzeeinwirkung, oder Chemikalien, geht in der Regel mit dem Verlust oder Beeinträchtigung seiner Funktion einher. So kann beispielsweise aus einem gekochten Hühnerei niemals mehr ein Küken schlüpfen. Falsch gefaltete Proteine sind oft die Ursache unterschiedlichster Krankheiten wie z.B.der zystischen Fibrose [7], der Progerie [8], BSE, von der Traberkrankheit (*Scrapie*), und der Creutzfeldt-Jakob-Krankheit [9].

Auf Basis der Kenntnis von Proteinstrukturen ist es heute möglich mit Hilfe von Computern gezielt Wirkstoffe gegen entsprechende Krankheiten zu entwickeln (*computer aided drug design*) [10]. Dies führt im weiterem zur Minimierung von Entwicklungskosten, wie auch der Verringerung der Anzahl an Tierversuchen [10]. Für die moderne Forschung, deren vorrangiges Ziel es ist biologische Vorgänge und Krankheiten insbesondere auf molekularer Ebene zu verstehen, ist die Kenntnis über den strukturellen Aufbau von Proteinen von fundamentaler Bedeutung. Diese Tatsache spiegelt sich besonders in der Initierung von ehrgeizigen Forschungsprojekten wieder, deren Ziel es ist, sämtliche vorhandene Proteinsequenzen innerhalb einer Zelle oder eines Organismus strukturell zu charakterisieren (*structural genomics*) [11]. Die mit dieser Entwicklung einhergehende stark zunehmende Anzahl strukturell zu charakterisierender Proteine (Abb. 1.3) stellt für die Strukturbiologen eine große Herausforderung dar.

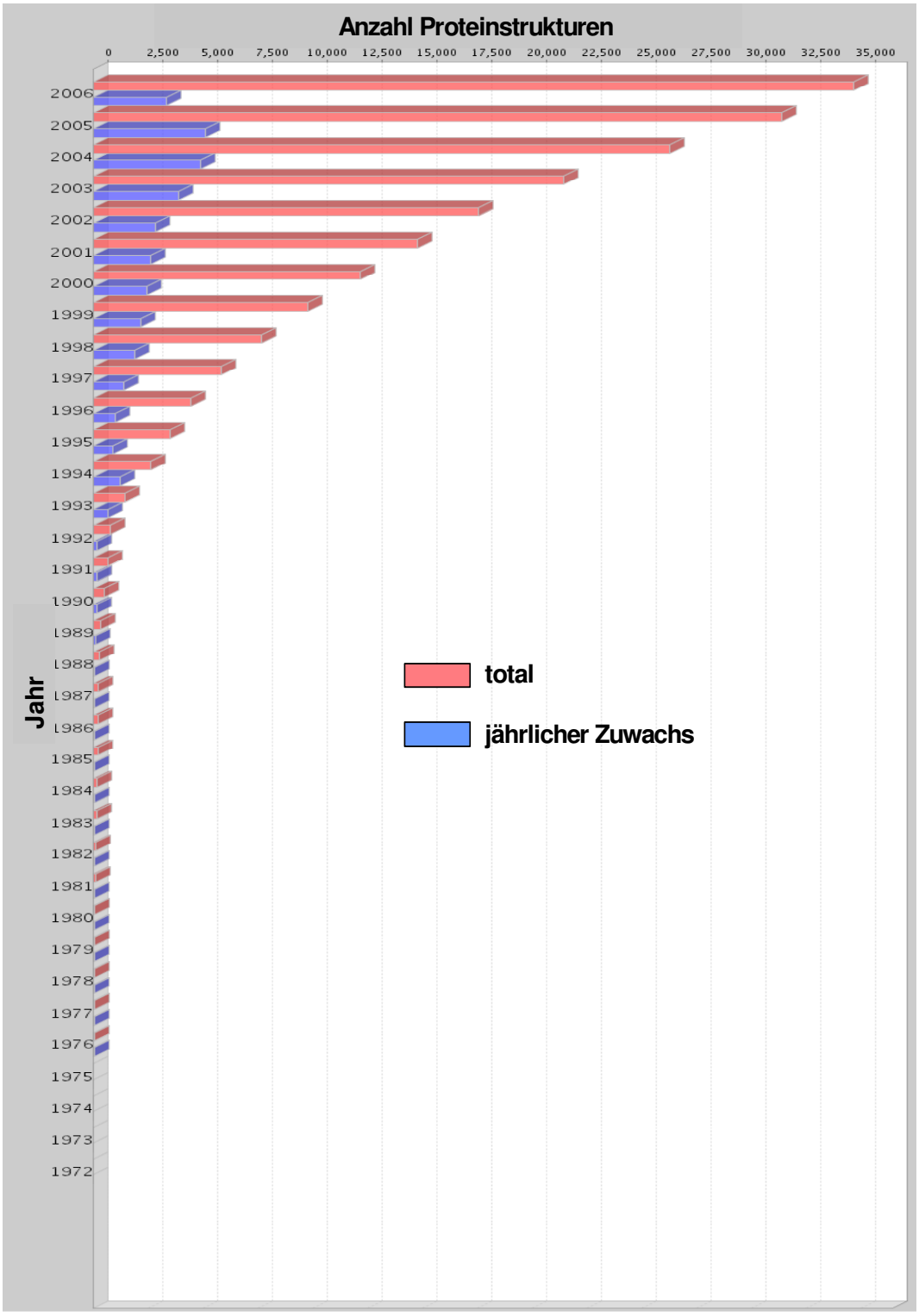


Abbildung 1.3: Zunahme der Anzahl aufgeklärter Proteinstrukturen [47]

1.2 Die NMR-Spektroskopie als Methode zur Proteinstrukturaufklärung

Zur Gewinnung von Informationen über die Struktur und Dynamik biologischer Makromoleküle, stehen dem Forscher heute zahlreiche unterschiedliche experimentelle sowie computergestützte Methoden zur Verfügung. Die Wahl der Methode hängt von unterschiedlichen Faktoren wie z.B. der erwünschten Genauigkeit und Vollständigkeit der Strukturinformationen über das zu untersuchende Molekül, der Erfahrung im Umgang mit entsprechenden Verfahren oder deren Verfügbarkeit, dem eingeplanten Arbeits- und Zeitaufwand und ähnlichem, ab.

Mit den meisten spektroskopischen Methoden wie z.B. der Infrarotspektroskopie [13], CD-Spektroskopie [15], Ramanspektroskopie [16], Floureszenzspektroskopie [17] oder Neutronenstreunung [14] lassen sich lediglich Teilaspekte über den strukturellen Aufbau biologischer Makromoleküle analysieren.

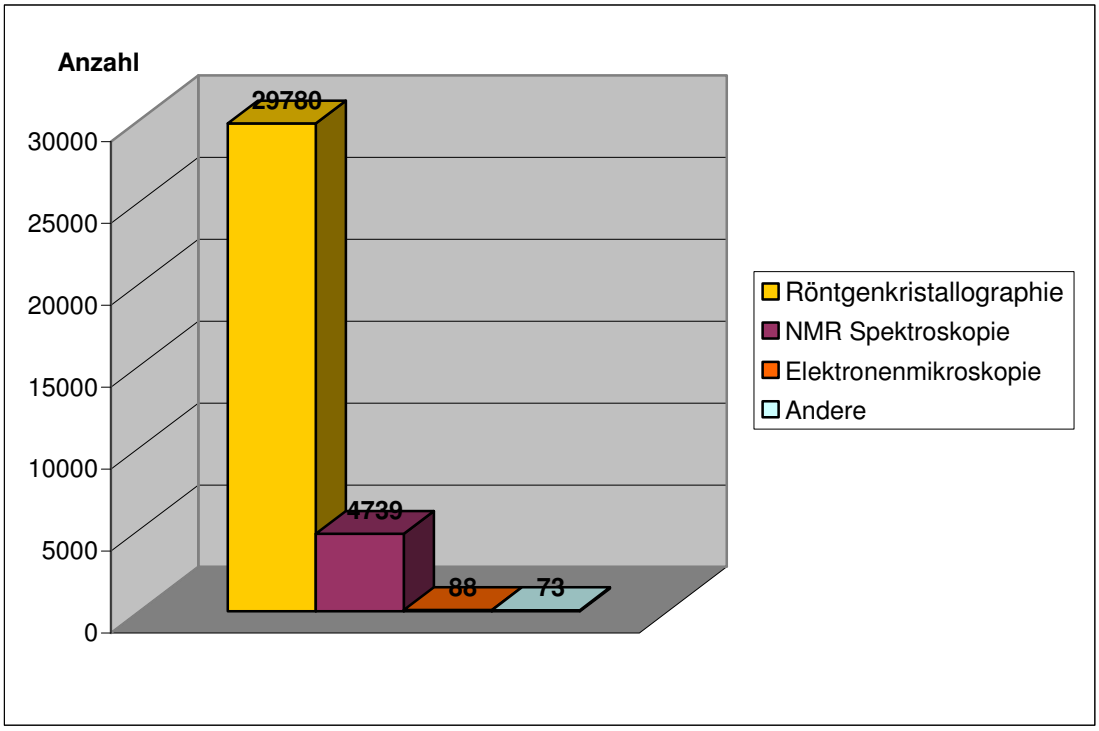


Abbildung 1.4: Anzahl abgelegter Proteinstrukturen innerhalb der PDB Datenbank [47]. Die Zahlen sind nach der angewandten Methode der Strukturaufklärung aufgeführt (Stand 31. 7. 2006).

Dazu zählen beispielsweise die Verteilung und Existenz von Sekundärstrukturen innerhalb des Moleküls, so wie die Dynamik und Konformation bestimmter Bereiche der Struktur. Versuche die Faltung eines Proteins nur auf Basis der Aminosäuresequenz vorherzusagen oder zu simulieren, wie z.B. durch *Homologiemodelling* [21] , *Threading* [19] oder sog. *Ab-initio-Methoden* [20] konnten, aufgrund der Komplexität der Sachverhalte und der noch unzureichenden Rechenleistung der heutigen Computer, bislang nur sehr unbefriedigende Ergebnisse liefern [13].

Die Elektronenmikroskopie, mit der man in der Lage ist, Strukturen mit nahezu atomarer Auflösung abzubilden, ist aufgrund der relativ starken Wechselwirkung des Elektronenstrahls mit der Materie und der Modifikation der Probe durch notwendige präparative Maßnahmen, für die Untersuchung biologischer Makromoleküle nur bedingt geeignet [13].

Als die heute wichtigsten Methoden zur vollständigen Bestimmung von Proteinstrukturen mit atomarer Auflösung haben sich die Röntgenkristallographie und die NMR-Spektroskopie erwiesen.

Der Großteil der bis dato aufgeklärten Proteinstrukturen ist mit Hilfe der Röntgen-kristallographie bestimmt worden (Abb. 1.4). Sie wurde erstmals 1959 am Myoglobin des Pottwals von John Kendrew angewandt [5]. Die Methode hat allerdings drei entscheidende Schwächen:

1. Nicht aus allen Proteinen lassen sich die, für diese Technik benötigten, Kristalle züchten.
2. Dynamische Prozesse innerhalb des Moleküls können nur bedingt charakterisiert werden.
3. Es besteht die Gefahr, dass die innerhalb eines Kristalls eingebundenen Proteine nicht die unter physiologischen Bedingungen vorherrschende Struktur einnehmen.

Die NMR-Spektroskopie konnte sich insbesondere aufgrund der bahnbrechenden Beiträge, wie der Fourier-Transformationsspektroskopie durch Richard Ernst und Einführung der grundlegenden Methoden durch Kurt Wüthrich Mitte der achtziger Jahre [22] [23], zu einer der wichtigsten Methoden zur Aufklärung von Proteinstrukturen etablieren (Abb. 1.4). Auf die Grundlagen der NMR-Spektroskopie soll hier nicht näher eingegangen werden, und es wird auf die entsprechende Literatur [28][29] verwiesen.

Da sich die zu untersuchende Probe bei der NMR-Spektroskopie in Lösung befindet, fallen die oben aufgeführten Probleme bei Anwendung der Röntgenkristallographie weg.

Die Kernprobleme der NMR-Spektroskopie liegen hauptsächlich in der Komplexität und Menge der auszuwertenden Daten, welche den zeitlimitierenden Faktor des

Proteinstrukturbestimmungsprozesses ausmachen. Anfangs war die Größe der zu untersuchenden Proteine bei Anwendung homonuklearer 2D-NMR-Spektren, wegen der starken Zunahme von Signalüberlappungen und Signalverbreiterung mit steigender Molekülgröße, auf etwa 10kDa beschränkt [13]. Im Zuge der fortlaufenden Verbesserung der Technik und Methoden, lassen sich heute Proteinstrukturen in der Größenordnung mit bis etwa 30 kDa relativ einfach bestimmen [13].

Zu den wesentlichen technischen Verbesserungen zählen insbesondere die Einführung der gepulsten Fourier-Transformationsspektroskopie (FT-NMR) [13], so wie die Entwicklung von NMR Spektrometern mit immer höheren Protonenresonanzfrequenzen von mittlerweile bis zu 950 MHz. Diese führten zu einer wesentlichen Steigerung der Empfindlichkeit der Messmethode, und damit zu einer höheren Auflösung der resultierenden Spektren.

Der Einsatz heteronuklearer so wie drei- oder vierdimensionaler NMR-Experimente seit Anfang der neunziger Jahre, ermöglichte eine weitgehende Trennung der Signale innerhalb der NMR-Spektren, und reduzierte somit die Anzahl störender Signalüberlappungen [24]. Durch die in neuer Zeit angewandten NMR-Experimente wie *TROSY* [26] und *CRINEPT* [27], ist es nun bereits auch möglich, durch Verringerung von Relaxationsverlusten gegenüber konventionellen Aufnahmetechniken, NMR-Spektren von Proteinen mit über 100 kDa auszuwerten [25].

Aufgrund der immer weiter steigenden Leistungsfähigkeit von Computern sowie der Neu- und Weiterentwicklung von Programmen zur Auswertung der komplexen und umfangreichen NMR-Messdaten, ist es in immer kürzerer Zeit möglich, die Struktur eines Proteins aufzuklären.

Den für die Auswertung der NMR-Messdaten bzw. für den Strukturbestimmungsprozess benötigten Arbeitschritten stehen heute mehrere unterschiedliche Programme zu Verfügung. Wesentliche Arbeitschritte beim Strukturbestimmungsprozess unter Anwendung der NMR-Spektroskopie sind die Prozessierung und Visualisierung der NMR Rohdaten (z.B. mit *XWINNMR* [31], *AZARA* [59], *TRIAD* [60]), Signal- und Multipletterkennung (z.B. mit *ATNOS* [32], *AUTOPSY* [48]), sequentielle Zuordnung (z.B. mit *CARMA* [49], *GARANT* [50], *MONTE* [51];), Automatische Zuordnung von NOESY Spektren (z.B. mit *ARIA* [34], *CANDID* [52], *SANE* [34], *KNOWNOE* [53];), Strukturrechnung (z.B. mit *AMBER* [54], *CNS* [55], *CYANA* [56];) und die Qualitätsbeurteilung von Strukturen (z.B. mit *PROCHECK_NMR* [35], *RFAC* [57], *AQUA* [35], *PROSA II* [58];).

Es gibt auch Softwarepakete wie z.B. *ANSIG* [36], *AURELIA* [37], *AUREMOL* [30] oder *FELIX* [38], welche zugleich mehrere der obigen Funktionen enthalten.

Kernziel der Entwicklung zukünftiger Programme ist vor allem die Automatisierung des Proteinstrukturbestimmungsprozesses auf der Basis der gegebenen NMR-Messdaten. Hierzu zählt vor allem die Zuordnung von NOESY-NMR-Spektren, welche zu den zeitaufwendigsten wie auch fehleranfälligsten Arbeitsschritten überhaupt zählt.

1.3 Zuordnung von NOESY-NMR-Spektren

Aus NOESY-NMR-Spektren gewinnt man den Großteil der für die Strukturbestimmung benötigten Information über interatomare Abstände. Die Zuordnung von NOESY-NMR Spektren stellt den zeitraufwendigsten Schritt bei der Proteinstrukturbestimmung mit Hilfe der NMR-Spektroskopie dar. Vor nicht allzu langer Zeit hat die Zuordnung der meist mehreren Tausend Signale oft mehrere Monate bis über ein Jahr in Anspruch genommen. Deshalb ist die Etablierung automatischer Zuordnungsmethoden von besonderer Wichtigkeit. Die Zuordnung von NOESY-NMR-Signalen beruht in der Regel auf der Basis bekannter chemischer Verschiebungen. Hauptproblem dabei ist, dass aufgrund von Signalüberlappungen innerhalb der Spektren, technischer Grenzen des erreichbaren Auflösungsvermögens oder unvollständiger sequentieller Zuordnung, oft ein Großteil der vorhandenen NOESY-Signale sich nicht eindeutig zu einem bestimmten Atompaar zuordnen lassen. Das Programm *KNOWNOE*, welches ein Teil des Programmpakets *AUREMOL* ist, verfolgt zur Lösung des Problems einen statistischen Ansatz. Dieser ermöglicht es, auf Basis struktureller Informationen, im Fall von zwei- oder drei Zuordnungsmöglichkeiten für ein NOESY-Signal, die jeweils Wahrscheinlichste zu berechnen. Die Berechnungsgrundlage bildeten hierbei Abstandshäufigkeitsverteilungen zwischen unterschiedlichen Atompaaren gewonnen aus einer Vielzahl strukturell bekannter Proteine.

Ziel des statistischen Berechnungsverfahren ist zu einem, möglichst viele unter den jeweils vorhandenen zwei- oder dreideutigen NOESY-Signalen zu finden, welche zu mindestens 90% von einem bestimmten Atompaar erklärt werden, und zum anderen, dem Signal das entsprechende signaldominierende Atompaar mit einer hohen Wahrscheinlichkeit zuzuweisen. Der statistische Ansatz hat den Vorteil, gegenüber anderen automatischen Zuordnungsverfahren, bereits vor der ersten Strukturrechnung Mehrdeutigkeiten aufzulösen. Dies verhindert in erster Linie die Einbeziehung von unrealistischen interatomaren Abständen in die Strukturrechung aufgrund falscher Zuordnungen. Bis jetzt lieferte der genannte statistische Ansatz noch recht unbefriedigende Ergebnisse. So war z.B. die Anzahl der

zugeordneten Signale relativ gering und zugleich ein hoher Anteil der erstellten Zuordnungen falsch war.

Wesentliches Ziel der Arbeit war die Zuordnungsqualität von zwei- und dreideutigen NOESY- Signalen bezüglich der Zuordnungsanzahl, wie auch der Zuordnungssicherheit zu verbessern. Dies sollte durch eine, im Vergleich zu Früher, qualitativ hochwertigere und stark erweiterte Datenbasis, auf die während der automatischen Zuordnung zugegriffen werden kann, erreicht werden. Dabei konnte auch eine Einschätzung über den statistisch relevanten Informationsgehalt bezüglich interatomarer Abstände innerhalb einer Vielzahl strukturell bekannter Proteine gewonnen werden.

2.0 Grundlagen

2.1 Das Programm *AUREMOL*

2.1.1 Allgemeines

AUREMOL ist ein Softwarepaket zur halbautomatisches Auswertung von NMR-Spektren zur Proteinstrukturbestimmung. Es wurde am Institut für Biophysik und physikalischer Biochemie der Universität Regensburg, in Zusammenarbeit mit der Firma Bruker Bio Spin, entwickelt und wird laufend verbessert und ausgebaut. Die wichtigsten Funktionen sind:

1. Interaktive Bearbeitung von 2D/3D NMR-Spektren.
2. Rückrechnung von NOESY-NMR-Spektren.
3. Berechnung von R-Faktoren [57].
4. Automatische Zuordnung von 2D/3D NOESY-NMR-Spektren.

Das Konzept von *AUREMOL* verfolgt einen molekülorientierten Ansatz (*top down* Strategie), der im Gegensatz zu herkömmlichen angewandten NMR-zentrierten Ansatz (*botton up* Strategie) steht [81]. Beim NMR-zentrierten Ansatz versucht man auf der Basis vollständig und korrekt zugeordneten NMR-Spektren die Proteinstruktur zu erhalten. Dieses Vorgehen erfordert allerdings eine hohe Anzahl von NMR-Experimenten. Das Ziel des molekülorientierten Ansatz hingegen ist es, mit möglichst wenig NMR-Experimenten und Unterstützung von mit bereits im Vorfeld gesammelten allgemeinen Daten über Proteine so wie zusätzlichen Informationen über das zu untersuchende Protein bzw. der NMR-Probe, die korrekte dreidimensionale Struktur zu bestimmen.
Als allgemeines bzw. von der spezifischen NMR-Probe unabhängiges Wissen gehören z.B. die chemische Struktur von Aminosäuren, Definition von verschiedenen NMR-Experimenten, statistische Erwartungswerte chemischer Verschiebungen und ihrer Verteilungen, J-Kopplungskonstanten, Karplusparameter und temperaturabhängige Viskositätskonstanten. Diese Daten sind in einer internen Datenbank von *AUREMOL* abgespeichert.
Zum NMR probenabhängigen Wissen zählen die Primärsequenz, das in der Probe befindliche Protein, die chemische Zusammensetzung der NMR Probe (z.B. Pufferzusammensetzung)

und die Bedingungen die während der Messung herrschten, wie Temperatur, Druck und pH-Wert.

2.1.2 Funktionalität der Programmkomponenten

Im folgendem sollen die Funktionen sowie die Zusammenarbeit der einzelnen Programmkomponenten des Softwarepakets *AUREMOL* während des Strukturbestimmungsprozesses anhand der Abbildung 2.1 erläutert werden.

Zunächst müssen die aufgenommenen NMR-Spektren vor der Benutzung mit *AUREMOL* vorprozessiert bzw. fouriertransformiert und gefiltert werden. Dies kann z.B. über das Programm *XWINNMR* geschehen (Abb. 2.1 *3*). Danach können die Spektren mit *AUREMOL* visualisiert so wie manuell oder automatisch bearbeitet werden (Abb. 2.1 *8*). Wichtige Funktionen zur Bearbeitung von NMR-Spektren sind z.B. die manuelle oder automatische Ermittlung von NMR-Signalen *(Peak Picking)*, Volumenintegration und automatisches Entfernen von Rauschen und Artefakten.

Bei der Strukturbestimmung unter *AUREMOL* handelt es sich um einen iterativen Prozess. Als Ausgangspunkt wird eine Modellstruktur (Abb. 2.1 *7*) benötigt. Diese kann man unter Einbeziehung von Informationen aus der eigenen internen Datenbank von *AUREMOL* (Abb. 2.1 *1*) und den molekülspezifischen Daten (Abb. 2.1 *3*) beispielsweise über Homologiemodelling mithilfe des Programmmoduls *PERMOL* [82] (Abb. 2.1 *5*) erhalten. Es ist aber auch möglich mit einer ausgesteckten Peptidkette als Startstruktur zu beginnen.

Über das Programmmodul *RELAX* [42] (Abb. 2.1 *11*) kann man nun aus einer gegebenen Startstruktur ein NMR-Spektrum zurückrechnen. Die benötigten Berechnungen basieren dabei auf der vollständigen Relaxationsmatrix Analyse. Mit *RELAX* kann man sowohl ^{1}H 2D –NOESY- wie auch ^{1}H, $^{15}N/^{1}H$, ^{13}C NOESY-HSQC NMR-Spektren simulieren. Durch Vergleich von simulierten und experimentellen NMR-Spektren werden fehlende chemische Verschiebungen in den Spektren ermittelt und zugeordnet [81](Abb. 2.1 *12*).

Wenn ein Großteil der chemischen Verschiebungen zugeordnet ist (sequentielle Zuordnung), kann man nun über das Programmmodul *KNOWNOE* vorhandene 2D/3D NOESY-NMR-Spektren des zu untersuchenden Proteins automatisch zuordnen lassen (Abb.2.1 *15*). Dabei dient die bereits vorhandene Modellstruktur zur Beschränkung von Zuordnungsmöglichkeiten.

Über das Programmmodul *REFINE* [83] lassen sich nun aus zugeordneten NOESY-Signalen Atomabstände berechnen. (Abb.2.1 *14*). Die Ermittlung der Atomanstände beruht dabei prinzipiell auf der iterativen Anpassung von simulierten NOESY-Signalen an ihre

entsprechenden Signale im experimentellen NOESY-Spektrum (s. Kap. 2.3). Die somit erhaltenen Atomabstände können nun zusammen mit bereits anderen vorhandenen Abstands- und Winkelbeschränkungen in die Strukturrechnung einbezogen werden (Abb. 2.1 *13*). Die Strukturrechnung ist kein Bestandteil des Softwarepakets *AUREMOL*. Diese kann beispielsweise über die Programme *CNS* oder *DYANA* [90] erfolgen. Im nächsten Schritt wird die bereits vorhandene Modellstruktur durch die, aus der Strukturrechnung erhaltene, Struktur ersetzt. Ausgehend von der jeweils neuen Modellstruktur, werden die beschriebenen Arbeitsschritte solange wiederholt, bis die gewünschte Güte der berechneten Poteinstruktur erreicht ist (Abb. 2.1 *9*). Nach erfolgter Strukturrechnung kann eine Qualitätsbeurteilung der erhaltenen Struktur mithilfe von R-Faktoren *(residual indicis)* vorgenommen werden. R-Werte oder R-Faktoren sind in diesen Zusammenhang ein Maß dafür inwieweit experimentelle und simulierte NMR-Spektren Daten übereinstimmen und somit eine Aussage über die Übereinstimmung der errechneten mit der wirklichen Struktur erlauben. Die Berechnung von R-Werten erfolgt über das Programmmodul *RFAC* (Abb. 2.1 *10*).

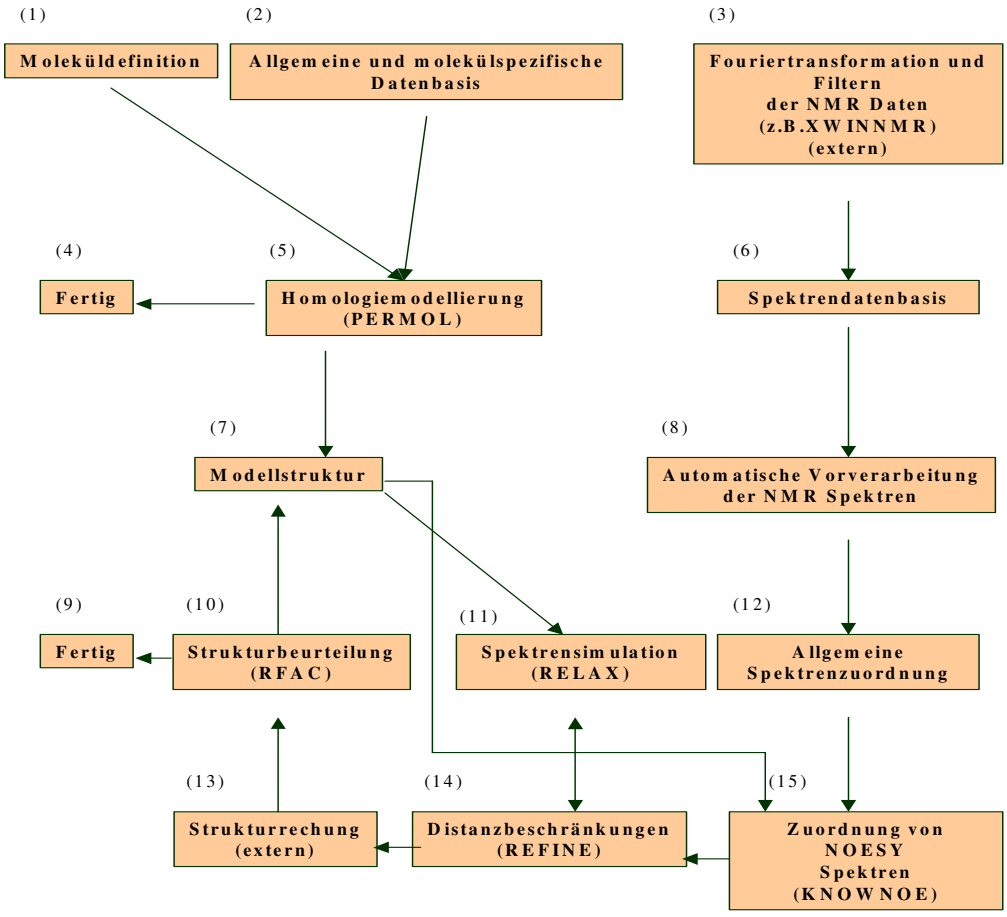


Abbildung 2.1: Das Konzept von *AUREMOL*. Die Grafik zeigt die wichtigsten Funktionen des Programms *AUREMOL* zusammen mit seinen Programmkomponenten. Der Informationsfluss zwischen den einzelnen Komponenten ist durch Pfeile verdeutlicht. Funktionen, die nicht zum Programmpaket gehören, sind mit dem Vermerk „extern“ gekennzeichnet.

2.2 Das NOESY Experiment

Das wichtigste Experiment bei der Proteinstrukturaufklärung mithilfe der NMR-Spektroskopie ist das NOESY-Experiment *(Nuclear Overhauser Effect Spectroscopy).*

Aus NOESY-NMR-Spektren werden vor allem die für die Aufklärung der räumlichen Struktur benötigten Informationen über interatomare Abstände (Wasserstoffatomkerne) gewonnen. Hierbei macht man sich den abstandsabhängigen *Kernoverhausereffekt* [29] zu Nutze. Dieser bewirkt eine nachweisbare Polarisationsänderung von räumlich benachbarten Atomkernen, welche über die Dipol-Dipol Wechselwirkung vermittelt wird.

Der grundlegende Ablauf eines 2D-NOESY-Experiments besteht aus einer Folge von jeweils drei hintereinander liegenden 90^o Pulsen (Abb. 2.2).

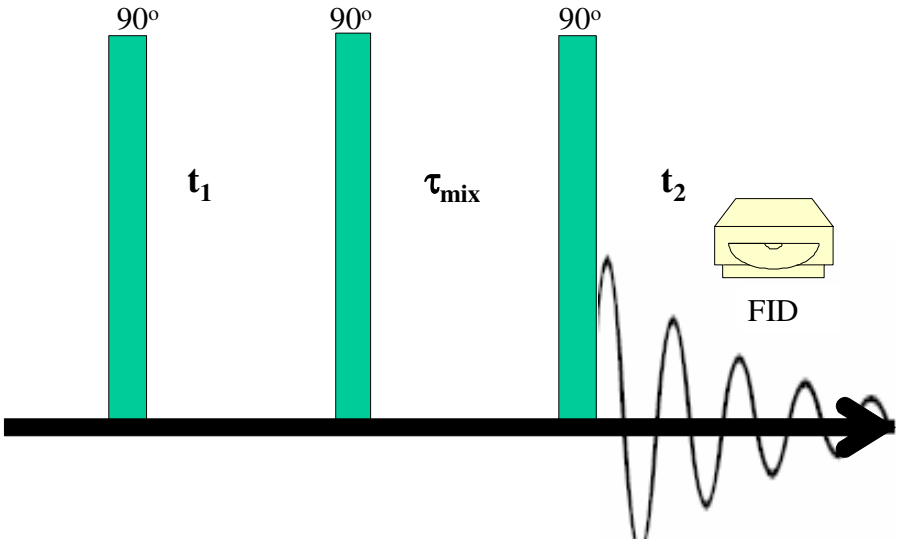


Abbildung 2.2: Ablauf des Standart 2D-NOESY-Experiments

Der erste Puls erzeugt zunächst Transversalmagnetisierung (x/y -Ebene). Die folgende variable *Evolutionszeit* t_1 dient zur Entwicklung der Spinsysteme, in der die Magnetisierung entsprechend der chemischen Verschiebung frei präzedieren kann.

Nach jeder Pulsfolge wird t_1 jeweils um einen festen Betrag Δt_1 erhöht, dessen Größe über das Nyquist -Theorem bestimmt wird:

$$\Delta t_1 = \frac{1}{2\upsilon_{\max}} \qquad (2.1)$$

ν_{max} entspricht hierbei der Spektrenweite in Hz.

Der nun folgende zweite 90° Puls konvertiert transversale (x/y-Ebene) in longitudinale (z-Achse) Magnetisierung. In der anschließenden *Mischzeit* τ_{mix} kann nun ein Austausch von Magnetisierung mittels Dipol-Dipol-Wechselwirkung zwischen den Atomkernen erfolgen. Um ein detektierbares Signal zu erhalten, wandelt ein dritter 90° Puls die vorhandene longitudinale Magnetisierung wieder in transversale Magnetisierung um. In der daraufhin folgenden *Detektionsphase* t_2 erfolgt nun die Aufzeichnung der Daten bzw. des FID (*Free Induction Decay,* Freier Induktionszerfall). Aus den aufgezeichneten Daten werden mittels Fourier-Transformation die t_2 und t_1-Zeitdomänen in entsprechende Frequenzdomänen ω_2 bzw. ω_1 umgewandelt.

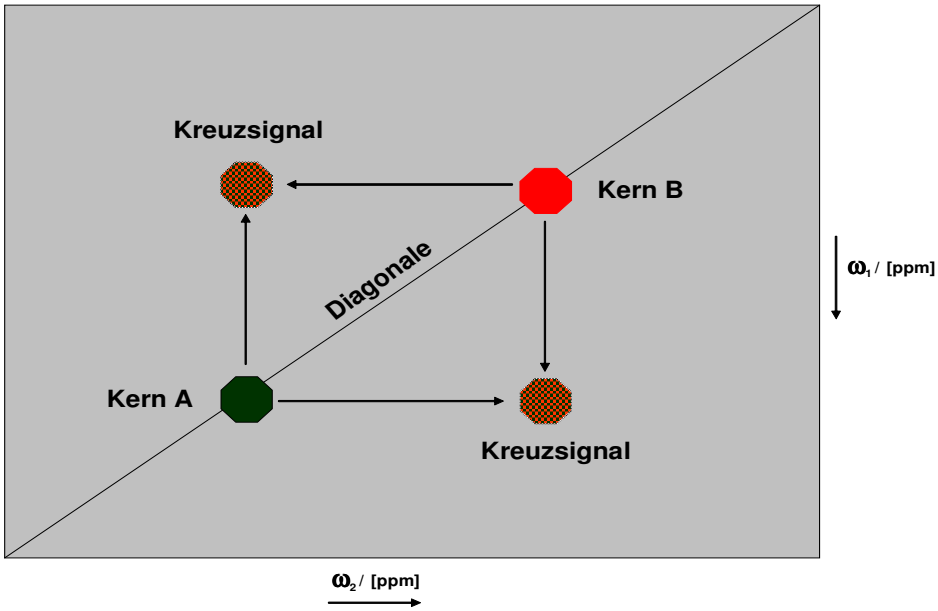


Abbildung 2.3: Aufbau eines 2D-NOESY-NMR-Spektrums (schematisch). Die Kreuzsignale zeigen jeweils die dipolare Kopplung bzw. räumliche Nähe der beiden Atomkerne A und B an.

Die resultieren Signale werden dabei in Form eines zweidimensionalen Spektrums mit entsprechenden Frequenzachsen für ω_2 und ω_1 visualisiert (Abb. 2.2). In der Regel lassen sich NOESY-Signale zwischen zwei Atomkernen mit einem Abstand von bis zu etwa maximal 0,5nm identifizieren. Zur Reduzierung von Rauscheffekten sowie Verbesserung des Auflösungsvermögens, werden die aufgezeichneten Daten vor der Fourier-Transformation meist durch speziell ausgewählte Funktionen modifiziert. Um vorhandene Signalüberlappungen zu trennen, können die 2D-NOESY-Pulsfolgen mit Pulsfolgen von heteronuklearen NMR-Experimenten wie z.B. dem 2D-HSQC-Experiment kombiniert werden.

Man erhält hierbei ein dreidimensionales NMR-Spektrum (^{13}C / ^{15}N–NOESY-HSQC) mit jeweils einer zusätzlichen dritten Frequenzachse für den Heterokern (^{15}N oder ^{13}C).

2.3 Berechnung interatomarer Abstände

In der Regel kommt bei der Berechnung von Atomabständen aus einem NOESY-Signal die sog. ISPA -Methode (*isolated spin-pair approximation*) zur Anwendung:

$$V_{ab} = \alpha r_{ab}^{-6} \tag{2.2}$$

V_{ab} ist hierbei das experimentelle Signalvolumen und r_{ab} der zu berechnende Abstand zwischen den Atomen *a* und *b*. Der Parameter α stellt einen proben- bzw. spektrenabhängigen Kalibrierungsfaktor dar. Diesen gewinnt man in der Praxis aus dem Volumen eines zugeordneten NOESY-Signals mit bereits bekanntem Abstand. Hierfür eignen sich NOESY-Signale von z.B. Wasserstoffatompaaren aus Sekundärstrukturen oder der gleichen Aminosäure, in denen meist bekannte charakteristische interatomare Abstände vorzufinden sind.

Ein Nachteil des Verfahrens ist, dass von der Gleichheit der Korrelationsfunktionen und Korrelationszeiten zwischen den unterschiedlichen Atomkernen ausgegangen wird. Außerdem wird von der Starrheit des Moleküls sowie der Richtungsunabhängigkeit der Rotationsdiffusion ausgegangen. Weiterer Nachteil ist die isolierte Betrachtung der beiden interessierenden Spins ohne den Einfluss andere Kerne zu berücksichtigen. Diese können beispielsweise zusätzliche Magnetisierung mittels Spindiffusion übertragen, was letztendlich zu einer Verfälschung des zu berechnenden Abstandes führen kann. Der Einfluss anderer Spins macht sich allerdings erst bei längeren Mischzeiten wesentlich bemerkbar. Die ISPA-Methode kann somit nur für verhältnismäßig kurze Mischzeiten brauchbare Ergebnisse liefern.

Unter Anwendung der Simulation von NOESY-Spektren ist, im Gegensatz zur ISPA-Methode, eine exaktere Abstandbestimmung, auch für längere Mischzeiten, aus experimentellen NOESY-Spektren möglich. Sie ist innerhalb von Spektrensimulations-programmen wie z.B. *IRMA* [84], *MARDIGRAS* [85], *MORASS* [86], *NO2DI* [87], *MIDGE* [88], ein Programm von Kim und Reid [89] sowie *RELAX* über das Programmmodul *REFINE* implementiert. Zentraler Aspekt ist hierbei die Berechnung der vollständigen Relaxationsmatrix [42]. Sie beschreibt die Übertragung der Magnetisierung während der Mischzeit eines NOESY-Experimentes. Hierbei werden prinzipiell alle dipolar gekoppelten Spins als Netzwerk betrachtet. In der Regel beschränkt man sich dabei allerdings auf

unmittelbar benachbarte Kerne. Die vorhandenen Programme unterscheiden sich hauptsächlich in der Anwendung bzw. Berücksichtigung der unterschiedlichen Bewegungsmodelle für Moleküle bei der Berechnung der Relaxationsmatrix voneinander.
Der Grundalgorithmus vom Programm *REFINE* basiert auf der iterativen Optimierung der Relaxationsmatrix durch Vergleich von experimentellen mit simulierten NOESY-Signalen. Dabei werden, ausgehend von einer Modellstruktur abgeleiteten initialen Relaxationsmatrix, die Relaxationsraten σ_{ij} der jeweils folgenden Iteration *n+1* aus den Raten der vorhergehenden Iteration $\sigma_{ij}(n)$ berechnet:

$$\sigma_{ij}(n+1) = \sigma_{ij}(n)\frac{\ln cA_{ij}(\exp)}{\ln A_{ij}(n, sim)} \tag{2.2}$$

A_{ij}(exp) stellt das experimentelle NOESY-Signalvolumen und $A_{ij}(n,sim)$ jeweils das simulierte NOESY-Signalvolumen beim n'ten Iterationsschritt für die beiden korrespondierenden Kerne *i* bzw. *j* dar. Mit der Variablen *c* sollen unbekannte technische und experimentelle Faktoren berücksichtigt werden.
Aus den jeweils erhaltenen Relaxationsraten werden nun wieder neue NOESY-Signale berechnet mit:

$$V_{ij}(\tau_{m,}\tau_r) = \alpha \cdot [\exp(-\tau_m \cdot R)]_{ij} \cdot \sum_k [1-\exp(-\tau_r \cdot R)]_{ik} \tag{2.3}$$

mit

$$\alpha = \frac{\sum A_{ij}^{ex} \cdot A_{ij}^{sim}}{\sum (A_{ij}^{sim})^2} \tag{2.4}$$

R ist die Relaxationsmatrix, welche die Relaxationsraten σ_{ij} enthält, τ_m ist die Mischzeit und τ_r der Zeitraum zwischen Beginn der Aufnahme des FID's und der nächsten Relaxationszeit.
Der Vorfaktor α sorgt für die Vergleichbarkeit von simulierten und experimentellen NOESY-Signalen. Die Folge aus Berechnungen von Relaxationsraten und NOESY-Signalen wird solange wiederholt, bis Formel 2.2 konvergiert ist bzw. sich experimentelle und simulierte Signalvolumen nicht mehr signifikant unterscheiden. Danach kann der Abstand d_{ij} der in Frage stehenden Kerne *i* und *j* über direkt aus der Relaxationsmatrix entnommen werden

2.4 Programme zur automatischen Zuordnung von NOESY-NMR-Spektren

Hier sollen einige die Grundkonzepte einer der gängigsten Programme zur automatischen Zuordnung von NOESY-NMR-Spektren kurz vorgestellt werden. Die meisten der vorhandenen Programme benötigten eine zuvor durchgeführte Zuordnung chemischer Verschiebungen (sequentielle Zuordnung) als Ausgangsinformation. Eine Ausnahme ist das Programm *CLOUDS* [98], welches während der Strukturrechnung eine Zusammenstellung von nur über NOE's miteinander verbundenen Wasserstoffatomen (engl. *cloud*) als Modell benutzt und im Allgemeinen nur für sehr kleine Moleküle zuverlässig funktioniert. Kernproblem bei der Zuordnung von NOESY-NMR-Spektren sind vor allem NOESY-Signale, die sich nicht eindeutig zu einem bestimmten Atompaar innerhalb des Proteins zuordnen lassen (mehrdeutige Signale). Dies kann folgende Ursachen haben:

1. Begrenztes Auflösungsvermögen des experimentellen NMR-Spektrums.

2. Die Zuordnung der chemischen Verschiebungen ist nicht komplett.

3. Überlagerung mehrerer Signale.

Die vorhandenen Algorithmen bzw. Programme unterscheiden sich hauptsächlich im Umgang mit mehrdeutigen NOESY-Signalen voneinander. Im folgendem werden die Grundprinzipien der Programme bzw. Algorithmen von *ARIA, SANE, NOAH, CANDID* und *AutoStructure* kurz erläutert.

Der Algorithmus von *ARIA* basiert im wesentlichem auf einer sich wiederholenden Kombination aus Resonanzzuordnungen und Strukturrechungen. Bei *ARIA* (*Ambiguous Restraints for Iterative Assignment*) werden mehrdeutige NOESY-Signale in die Strukturrechung integriert. *ARIA* ist gekoppelt an die Strukturrechungsprogramme *X-PLOR* und *CNS.*

Der *SANE* (*Structure Assisted NOE Evaluation*) Algorithmus funktioniert ähnlich wie *ARIA*. Das Programm ist an die MD-Programme *DYANA* und *AMBER* gekoppelt. Es integriert, ähnlich wie *ARIA*, mehrdeutige Abstandsbeschränkungen innerhalb eines iterativen Prozesses von NOESY-Signalzuordnungen und Strukturrechungen.

Bei *NOAH* [92] handelt es sich um ein iteratives Verfahren, welches eine Kombination aus automatischer NOESY-Signalzuordnung, Strukturrechung und Analyse von Abstandsverletzungen benutzt, um die endgültige Struktur eines Proteins zu bestimmen. Das

Programm ist innerhalb der Distanzgeometrieprogramme *DIANA* [91] und *DIAMOD* [92] implementiert.

Bei dem Algorithmus von *CANDID* (*Combined automated NOE assignment and structure determination module*) handelt es sich um einen iterativen Ansatz zur automatischen Zuordnung von NOESY-Signalen und automatischer Erzeugung von 3D-Proteinstrukturen.

Er kombiniert Methoden aus *ARIA* und *NOAH* wie z.B. die Integration mehrdeutiger NOESY-Signale und die Benutzung von Zuordnungsfilter basierend auf einer bereits vorhandenen dreidimensionalen Modellstruktur. Zur Minimierung von Artefakten und Rauschen wendet das Programm Methoden wie das sog. *NetworkAnchoring* [93] und die Kombination von Abstandsbeschränkungen(*Constraint Combination)* [93] an.

Bei dem Programm *AutoStructure* [96] handelt es sich um ein Expertensystem, welches die gleichen Regeln zur Bestimmung von Abstandsbeschränkungen aus experimentellen NMR-Spektren anwendet wie ein menschlicher Experte. Der Ansatz zeichnet sich durch die Anwendung der Graphentheorie [44] zur Formulierung des Problems der Interpretation von NOESY-Signalen aus. Das Programm wendet zu Interpretation von NOESY-Signalen einen *„botton up“* topologiebeschränkten Abstandsnetzwerkalgorithmus an und erzeugt, zusammen mit den Strukturrechungsprogrammen *XPLOR, CNS* oder *DYANA*, automatisch 3D-Proteinstrukturen.

2.5 Das Programm *KNOWNOE*

2.5.1 Überblick

Im Rahmen der Arbeit durchgeführten Testreihen von automatischen NOESY-Signal-zuordnungen wurden mit dem Programm *KNOWNOE* durchgeführt. Deshalb wird auf dieses Programm näher eingegangen.

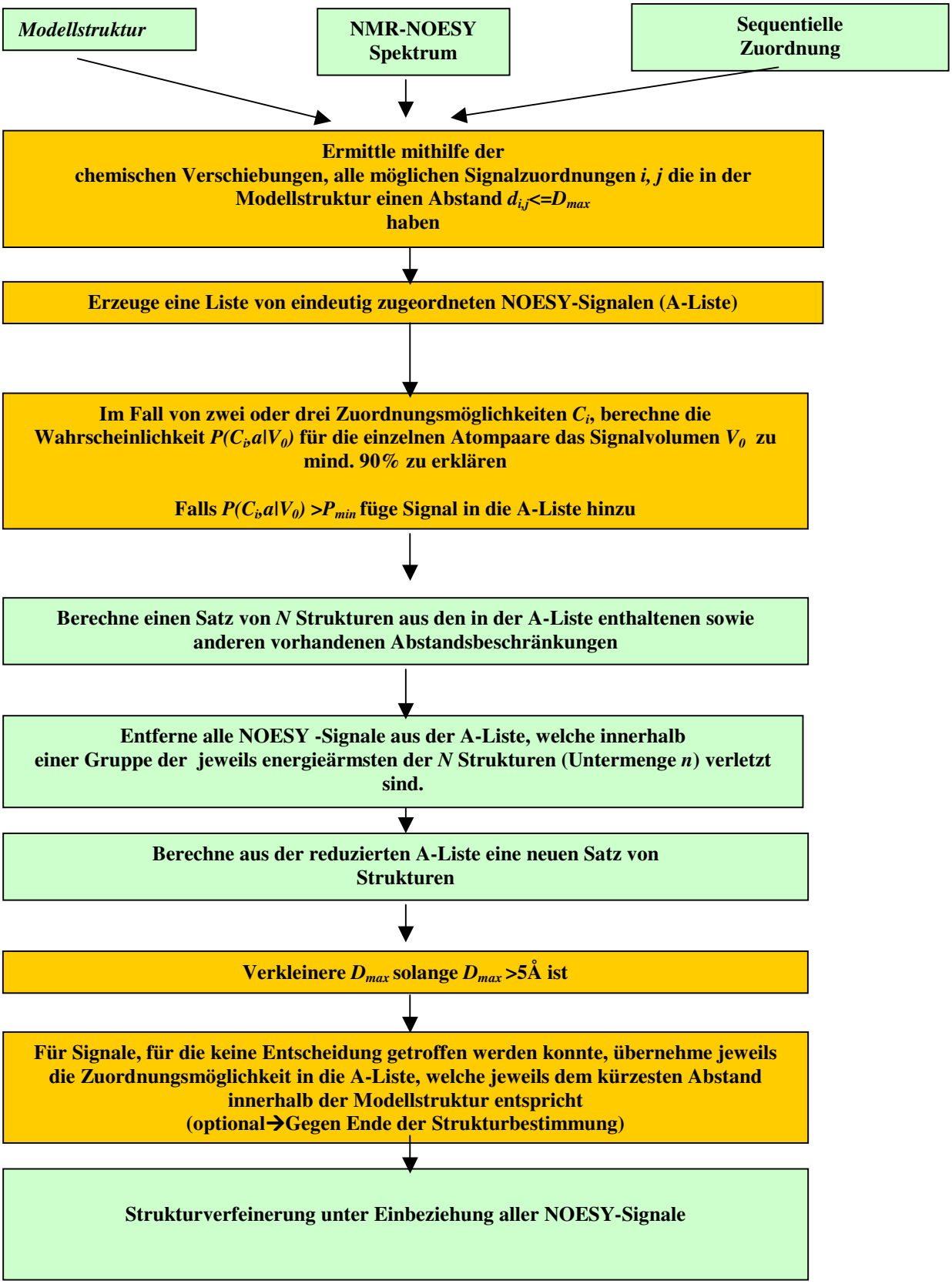
Hauptaufgabe des Programms *KNOWNOE* ist die automatische Zuordnung von 2D/3D-NOESY-NMR-Spektren. Das Programm ist ein wesentlicher Bestandteil des Programmpaketes *AUREMOL.* Ähnlich wie bei den meisten Programmen, erfolgt die Zuordnung von NOESY-Signalen bei *KNOWNOE* auf Basis bekannter chemischer Verschiebungen. Wesentlicher Vorteil von *KNOWNOE* ist, dass bereits vor der Strukturrechnung Mehrdeutigkeiten bei der Signalzuordnung aufgelöst werden können.

Dies ist besonders am Anfang der Strukturbestimmung wichtig, da falsche Zuordnungen hierbei die zu untersuchende Struktur in eine völlig falsche Konfomation zwingen können. Bei der Zuordnung von mehrdeutigen NOESY-Signalen verfolgt *KNOWNOE* einen

wissensbasierten Ansatz, welcher auf der Kenntnis interatomarer Abstände strukturell bekannter Proteine beruht. Hierbei versucht der Algorithmus zu einem solche Signale zu finden, deren Volumen von einem bestimmten Atompaar zu mehr als 90% erklärt werden und zum anderen den Signalen die entsprechenden Atompaare mit einer hohen Wahrscheinlichkeit zuzuweisen. Abbildung 2.4 zeigt den Algorithmus von *KNOWNOE* und seine Integration in den iterativen Prozess der Strukturbestimmung. Auf die in der Abbildung 2.4 aufgeführten Punkte wird im folgendem noch näher eingegangen werden.
Aus Abbildung 2.1 ist zu entnehmen, mit welchen anderen Programmmodulen des Softwarepakets *AUREMOL* das Programm *KNOWNOE* im unmittelbaren funktionalen Zusammenhang steht.

2.5.2 Signalzuordnungen aufgrund chemischer Verschiebungen

Jedes NMR aktive Atom im einem Protein lässt sich theoretisch aufgrund seiner charakteristischen chemische Verschiebung identifizieren. Sind alle oder ein Großteil der chemischen Verschiebungen von den Protonen des zu untersuchenden Proteins bekannt (sequentielle Zuordnung), ist es möglich aufgrund dieser Informationen von dem jeweiligen Protein aufgenommene NOESY-Spektren zuzuordnen.
Im ersten Arbeitschritt vergleicht das Programm *KNOWNOE* zunächst alle chemischen Verschiebungen aus der sequentiellen Zuordnung mit jeweils zwei (bei 2D-NOESY-NMR-Spektren) bzw. drei (bei 3D-NOESY-NMR-Spektren) der chemischen Verschiebungen eines bestimmten NOESY-Signals innerhalb des Spektrums. Falls dabei die Differenz einen bestimmten vom Benutzer vorgegebene Toleranzwert nicht überschreitet, wird der entsprechende in der sequentiellen Zuordnung stehende Atomname für das NOESY-Signal bzw. in die Zuordnungsliste (*Peakliste*) übernommen.



2.4 Algorithmus vom Programm *KNOWNOE* (übernommen aus [53])

Der benutzerdefinierte Toleranzwert *TOL(ppm)* für die chemischen Verschiebung hat den Zweck, vorhandene Unterschiede zwischen identischen Protonen, wie sie zwischen verschiedenen NMR-Experimenten z.B. aufgrund von Messungenauigkeiten auftreten, zu berücksichtigen. Die beschriebene Vorgehensweise ist in Abbildung 2.5 noch einmal verdeutlicht.

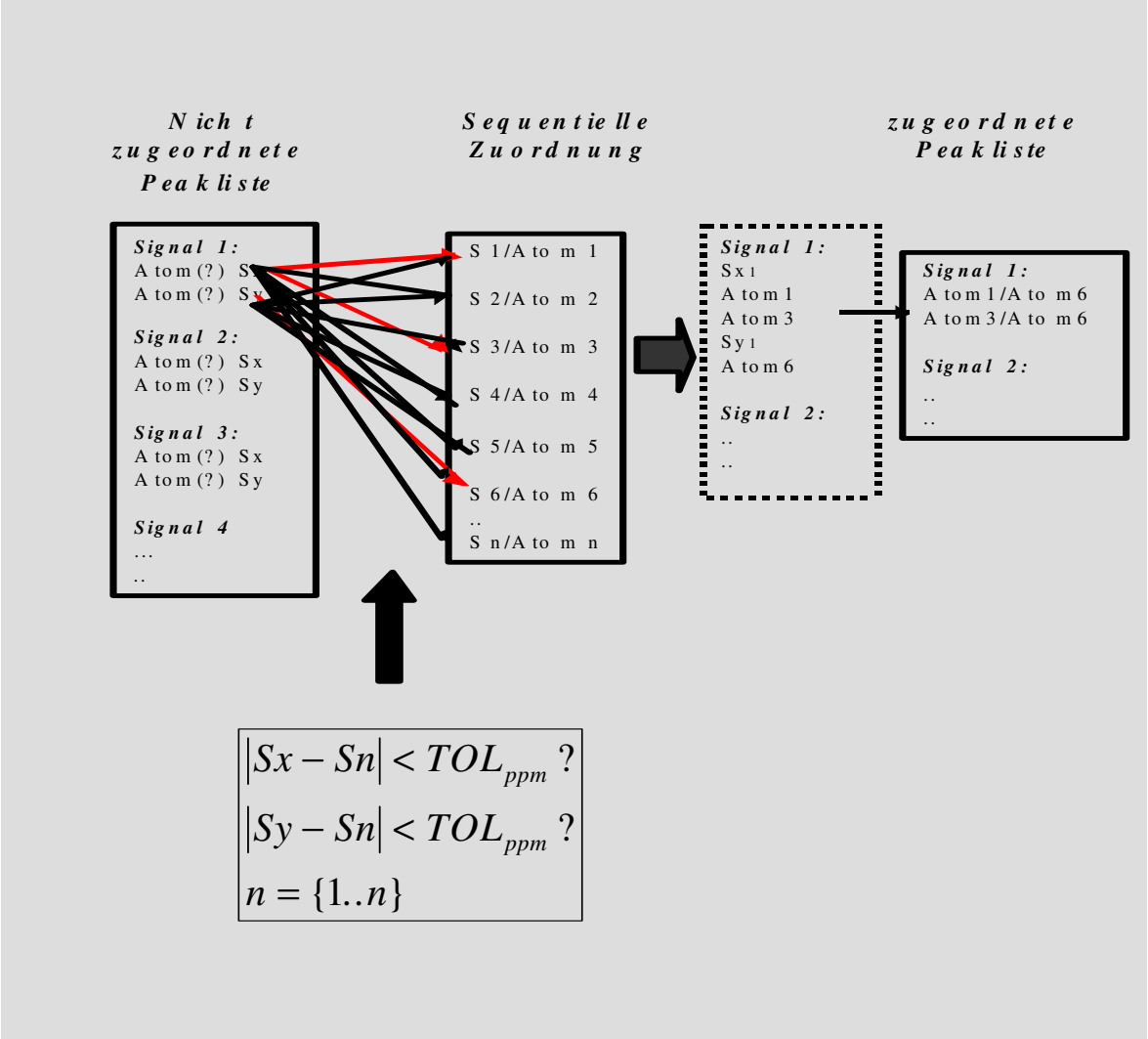


Abbildung 2.5: Zuordnung von NOESY-Signalen aufgrund chemischer Verschiebungen. Die experimentell ermittelten chemischen Verschiebungen S_x bzw. S_y der beiden Frequenzdomänen z.B. eines 2D-NOESY-NMR-Signals werden mit allen eingetragenen chemischen Verschiebungen S_1-S_n aus der Liste der sequentiellen Zuordnung verglichen. Ist die Differenz der chemischen Verschiebungen kleiner als der vorgegebene Toleranzwert TOL_{ppm}, wird das in der sequentiellen Zuordnung entsprechende Atom der zugehörigen chemischen Verschiebungen S_x bzw. S_y des bereffenden NOESY-Signals zugewiesen (rote Pfeile).

2.5.3 Behandlung mehrdeutiger NOESY-Signale

Wie bereits erwähnt, gibt es nach Zuordnung des NOESY-NMR-Spektrums mithilfe chemischer Verschiebungen, meist sehr viele Signale die mehr als nur eine Zuordnungsmöglichkeit erhalten haben. Zunächst werden diejenigen Zuordnungen wieder entfernt, deren Abstand innerhalb der gegebenen Modellstruktur einen vom Benutzer definierten Wert überschreiten. Zusätzlich können vorhandene Zuordnungsmöglichkeiten für ein NOESY-Signal durch Anwendung von *NetworkAnchoring* ausgeschlossen werden. Das Konzept des *NetworkAnchoring* beruht auf der Absicherung von NOESY-Signalen durch andere NOESY-Signale zwischen benachbarten Atomen. Trotz anfänglicher Reduzierung der vorhandenen Zuordnungsmöglichkeiten, bleiben in der Regel meist eine Vielzahl mehrdeutiger NOESY-Signale übrig. Im Fall von zwei- oder drei Zuordnungsmöglichkeiten, berechnet das Programm *KNOWNOE* die Wahrscheinlichkeiten für die jeweiligen Zuordnungsmöglichkeiten mindestens 90% des Signalvolumens zu erklären. Erreicht dabei eine Wahrscheinlichkeit einen bestimmten vom Benutzer definierten Mindestwertwert, wird die entsprechende Zuordnung dem betreffenden NOESY-Signal zugewiesen und in die Zuordnungsliste der eindeutig zugeordneten NOESY-Signale *(Peakliste*) übernommen.
Die Berechnung der Wahrscheinlichkeit für die gegebenen Zuordnungsmöglichkeiten erfolgt hierbei über Anwendung das *Bayes'schen* Theorems:

$$P(C_i, a|V_0) = \frac{P(C_i, a)\,P(V_0|C_i, a)}{\sum_{i=1}^{Nab} P(C_i, a)\,P(V_0|C_i, a)} \tag{2.11}$$

$P(C_i,a|V_0)$ ist die Wahrscheinlichkeit mit der der Anteil a vom Signalvolumen V_0 durch eine bestimmte Zuordnungsklasse (Atompaar) C_i erklärt wird. In dieser Anwendung beträgt a=0,9. N_{ab} entspricht der Anzahl der Zuordnungsmöglichkeiten für ein bestimmtes NOESY-Signal. $P(C_i,a)$ ist die *a priory* Wahrscheinlichkeit. Sie ist die Wahrscheinlichkeit, dass ein beliebiges Signalvolumen zu einer Zuordnungsklasse C_i gehört. Im einfachsten Fall, bei dem nur eine Zuordnungsmöglichkeit existiert, ist $P(C_1,a)=1$ und $P(C_i,a,i>1)=0$. Damit wird

$$P(C_i,a|V_0)=1 \tag{2.12}$$

Im Falle von genau zwei Zuordnungsmöglichkeiten, müssen vor der Berechnung von $P(C_i,a|V_0)$ zuerst die Wahrscheinlichkeiten $P(C_i,a)$ und $P(V_0|C_i,a)$ berechnet werden. Falls

keine andere Zuordnungsmöglichkeit in Betracht kommt, gilt für die *a priori* Wahrscheinlichkeiten für $i>2$:

$$P(C_i, a) = 0 \tag{2.13}$$

Falls nun die beiden Zuordnungsklassen für das gefragte Volumen V_0 dieselbe Wahrscheinlichkeitsverteilung besitzen, können die *a priori* Wahrscheinlichkeiten für $i=1$ und $i=2$ angenähert werden durch:

$$P(C_1, a) = P(C_2, a) = 0.5c_s \quad \text{mit } 0 \le c_s \le 1 \tag{2.14}$$

Bei c_s handelt es sich um ein Normalisierungskonstante, welche von der Form der Wahrscheinlichkeitsverteilung abhängt und sich bei der Berechnung von $P(C_i,a|V_0)$ herauskürzt.

Allgemein lassen sich die *a priori* Wahrscheinlichkeiten berechnen durch:

$$P(C_1, a) = \int_{V_0=0}^{\infty} \int_{V_1=aV_0}^{V_0} p_1(V_1)\, p_2(V_0 - V_1)\, dV_1 dV_0 \tag{2.15}$$

$$P(C_2, a) = \int_{V_0=0}^{\infty} \int_{V_2=aV_0}^{V_0} p_1(V_0 - V_2)\, p_2(V_2)\, dV_2 dV_0 \tag{2.16}$$

$p_1(V)$ und $p_2(V)$ sind die normalisierten Wahrscheinlichkeitsdichten ein gegebenes Volumen V mit der Zuordnung C_1 bzw. C_2 zu finden. Die Werte für $P(V_0|C_i,a)$ lassen sich für zwei Zuordnungsmöglichkeiten berechnen durch:

$$P(V_0 \mid C_1, a) = \int_{V_1=aV_0}^{V_0} p_1(V_1)\, p_2(V_0 - V_1)\, dV_1 \tag{2.17}$$

$$P(V_0 \mid C_2, a) = \int_{V_2=aV_0}^{V_0} p_1(V_0 - V_2)\, p_2(V_2)\, dV_2 \tag{2.18}$$

Im Falle von genau drei Zuordnungsmöglichkeiten können die *a priori* Wahrscheinlichkeiten analog nach den oben genannten Formeln gebildet werden:

$$P(C_i,a) = 0 \qquad (i>3) \tag{2.19}$$

$$P(C_i,a) = c_c/3 \qquad 0 \le c_s \le 1 \quad \text{für} \quad 1 \le i \le 3 \tag{2.20}$$

$$P(V_0|C_1,a) = \int_{V_1=aV_0}^{V_0} \int_{V_2=0}^{V_0-V_1} p_1(V_1)\,p_2(V_2)\,p_3(V_0-V_1-V_2)\,dV_2\,dV_1 \tag{2.21}$$

$$P(V_0 \mid C_2,a) = \int_{V_2=aV_0}^{V_0} \int_{V_1=0}^{V_0-V_2} p_1(V_1)\,p_2(V_2)\,p_3(V_0-V_1-V_2)\,dV_1\,dV_2 \tag{2.22}$$

$P(V_0|C_i,a)$ erhält man durch:

Die für die Berechnung der Integrale benötigten Wahrscheinlichkeitsdichten werden hierbei

$$P(V_0|C_3,a) = \int_{V_3=aV_0}^{V_0} \int_{V_1=0}^{V_0-V_2} p_1(V_1)\,p_2(V_0-V_1-V_3)\,p_3(V_3)\,dV_1\,dV_3 \tag{2.23}$$

aus Wahrscheinlichkeitsdichteverteilungen, welche auf der Basis großer Mengen bekannter interatomarer Abstände erzeugt worden sind (Kap. 4.1.5), bezogen. Bei mehr als drei Zuordnungsmöglichkeiten wird das NOESY-Signal nicht verwendet. Wenn am Ende des iterativen Strukturbestimmungsprozesses keine Entscheidung aufgrund der Wahrscheinlichkeit für ein mehrdeutiges NOESY-Signal getroffen werden kann, wird das Signalvolumen, in Abhängigkeit des jeweiligen Abstandes der in Frage stehenden Atompaare, innerhalb der Modellstruktur aufgeteilt.

2.5.4 Eingabeparameter zum Start von *KNOWNOE*

Hier soll gezeigt werden, welche Eingangsinformationen das Programm *KNOWNOE* zur seiner Ausführung benötigt.

In Abbildung 2.6 ist der gemeinsame Eingabedialog für die Programme *KNOWNOE* und *RELAX (Backcalculation parameters)* dargestellt. Bevor man den Eingabedialog öffnen kann, muss ein NMR-Spektrum geladen werden. Dies kann z.B. über den Menüpunkt *open*

spectrum im Menü *File* über der Arbeitsoberfläche von *AUREMOL* geschehen. Im folgendem werden die für den Start des Programms *KNOWNOE* benötigten Eingabeparameter erläutert:

1. *Input compound file*

In diesem Eingabefeld muss der Name für eine *compound*-Datei angegeben werden. Sie ist ein Bestandteil der Datenstruktur von *AUREMOL*. Hierbei handelt es sich um eine Datei, welche von der spezifischen NMR Probe unabhängige Informationen enthält.

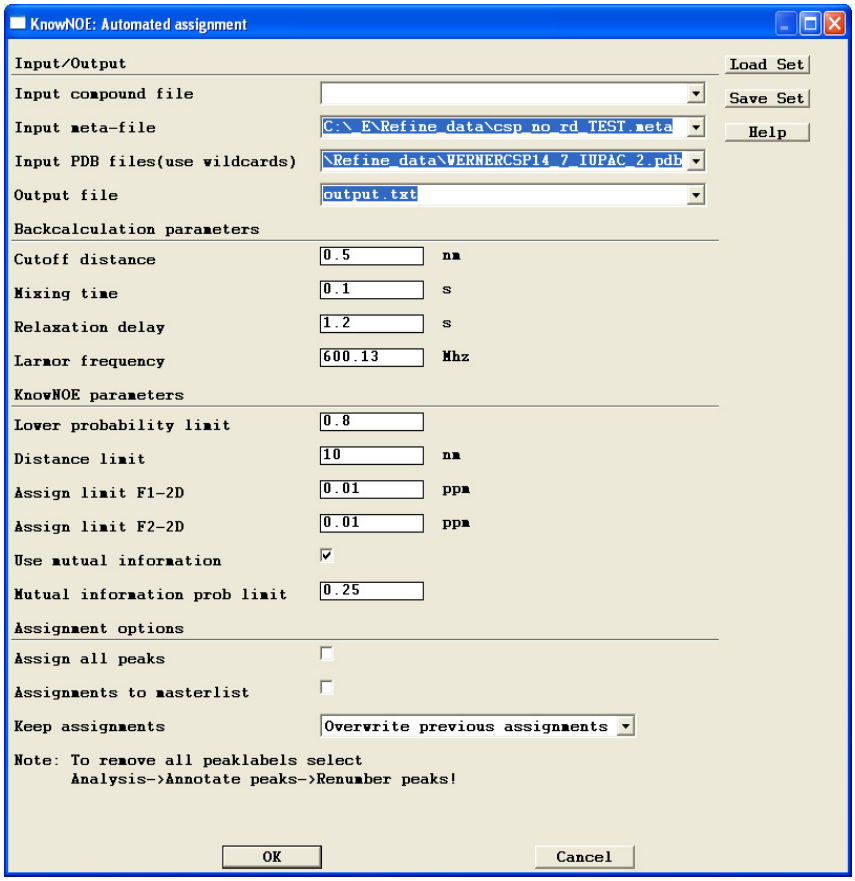


Abbildung 2.6: Eingabedialog für die Programme *KNOWNOE* und *RELAX*

Die *compound*-Datei enthält keine Angaben wie z.B. chemische Verschiebungen von Atomen, welche unter verschiedenen Versuchsbedingungen voneinander abweichen können. Die *compound*-Datei besteht aus drei Abschnitten, von denen nur der erste für das Programm *KNOWNOE* relevante Daten enthält. In diesen sind alle Atome des zu untersuchenden Proteins, in sequentieller Reihenfolge, spezifiziert. Abbildung 2.7 zeigt einen Ausschnitt einer *compound*-Datei. In der ersten Spalte ist die Position der Aminosäuren innerhalb der Sequenz angegeben. Als nächstes folgt die Zuordnungsnummer der Atome innerhalb der

Aminosäure, der Aminosäurename im Dreibuchstabencode, der Atomname (nach IUPAC), der Atomtyp, ein eventuell vorhandener Aliasname[1] und ein zu dem Atom magnetisch äquivalenter Kern (falls vorhanden) sowie Anisotropiewerte. In den übrigen Abschnitten befinden sich Informationen über dihedrale Winkel, kovalente Atombindungen, Karpluskonstanten und J-Kopplungskonstanten

```
COMPOUND: csp
section_sequenzdefinition
_Residue_seq_code
_Atom_num_code
_Residue_label
_Atom_name
_Atom_type
_Atom_alias
_Atom_equivalent
_Atom_CSA
1   1 MET   HN H   -     -          8.95
1   2 MET    N N   -     -        157.00
1   3 MET   CA C   -     -         40.00
1   4 MET   HA H   -     -          8.95
1   5 MET    C C   -     -        40.00
1   6 MET    O O   -     -       -
1   7 MET   CB C   -     -         40.00
…
66  14 GLU  OE1 O   -     -       -
66  15 GLU  OE2 O   -     -       -
66  16 GLU  HE2 H   -     -          8.95
end_section

section_bond
_Bond_start
_Bond_atom1
..
```

Abbildung 2.7: Ausschnitt aus einer *compound*-Datei

2. *Input meta-file*

Als nächstes wird die Angabe einer *Meta*-Datei verlangt, welche ebenfalls ein Bestandteil der Datenstruktur von *AUREMOL* ist [81]. Diese enthält Angaben über die einzelnen Bestandteile der NMR Probe wie z.B. ihrer Konzentration und eventuelle Isotopenmarkierungen, Daten über die während der Messung herrschenden Bedingungen wie z.B. pH-Wert und Temperatur. Weiter sind die Dateinamen aller vorhandenen *Compound*-Dateien aufgeführt. Als für das Programm *KNOWNOE* relevanten Informationen der *Meta*-Datei sind die Zuordnungen der einzelnen Atome zu chemischen Verschiebungen (sequentielle Zuordnung) zu nennen.

In Abbildung 2.8 ist ein Ausschnitt aus einer *Meta*-Datei dargestellt, welcher den Abschnitt für die sequentielle Zuordnung zeigt. Die Spalten, von links nach rechts, enthalten folgende

[1] Beispielsweise werden die Wasserstoffatome HB1, HB2 und HB3 einer Methylgruppe oft zu dem Aliasatom HB nach der IUPAC-Konvention zusammengefasst

Informationen: Die Aminosäurenummer, Atomnummerncode und ein Aliasname[1], wie sie für das Atom in der entsprechenden *Compoundatei* stehen. Als nächstes ist die chemische Verschiebung mit der zugehörigen Fehlerabweichung für das betreffende Atom angegeben.

```
SHIFTS:
_Residue_seq_code
_Atom_num_code
_Atom_alias
_Chem_shift_value
_Chem_shift_value_error
_Chem_shift_ambiguity_code
_Atom_class
_Linewidth
1   4    -          4.030  0.20   2   0   -
1   8    HB         1.601  0.20   2   0   -
1  11    -          2.189  0.20   2   0   -
1  12    -          2.019  0.20   2   0   -
1  15    HE         2.087  0.20   2   1   -
2   1    -          7.680  0.20   2   3   -
…
..
66   8    -         2.076  0.20   2   0   -
66   9    -         1.906  0.20   2   0   -
66  11    HG        2.189  0.20   2   0   -
END_SHIFTS
J_COUPL:
```

Abbildung 2.8: Ausschnitt aus einer *Meta*-Datei

Der folgende Wert gibt Aufschluss über die Sicherheit der Zuordnung. Zum Schluss wird die Zugehörigkeit zu einer in den oberen Abschnitten (hier nicht gezeigt) der *Meta*-Datei definierten Atomklasse und eine eventuell ermittelte Linienbreite angegeben

3. *Input PDB file*

Im nächsten Eingabefeld muss der Name bzw. Dateipfad einer PDB-Datei angegeben werden. Sie beinhaltet, neben Informationen wie z.B. die jeweils angewandten Methode der Strukturaufklärung, vorherrschende Versuchsbedingungen, Herkunft des betreffenden Proteins, Autor usw., vor allem die räumlichen Koordinaten der Atome des betreffenden Moleküls.

4. Output file

Im Eingabefeld *Output file* gibt der Benutzer den Namen einer Datei an in der alle automatisch zugeordneten NOESY-Signale vermerkt werden.

```
********* Name Nr Name Nr  Dist   Shift   Shift   Wahr  Moeg   Vol  Num_Wahr Struk_dist
peak assign HD1 81 HD2 81   2.406  -0.275   0.434 1.000000 1 326223 0   0.000
peak assign HD1 81 HD1 61   2.694  -0.275   0.608 1.000000 1 165794 0   4.019
peak assign HD1 81 HG2 77   2.358  -0.275   0.658 1.000000 1 368219 0   3.037
```

Abbildung 2.9: Ausschnitt aus der Ausgabedatei von *KNOWNOE*

Abbildung 2.9 zeigt einen Ausschnitt aus einer solchen Datei nach der Zuordnung eines 2D-NOESY-Spektrums. In jeder Zeile stehen Informationen über ein bestimmtes zugeordnetes NOESY-Signal. Folgende Daten, von links nach rechts, sind in jeweils einer Zeile gespeichert: Das erste und zweite Atom der Zuordnung mit der Sequenzposition der entsprechenden Aminosäure. Dann erfolgt eine Angabe über den räumlichen Abstand (Å) der beiden Atome, welcher aus dem experimentellen Volumens des Signals ermittelt wird (s. Kap.2.3). In den folgenden beiden Spalten stehen jeweils die chemischen Verschiebungen der Atome. Dann folgt eine Angabe über die Wahrscheinlichkeit, die für die betreffende Zuordnung berechnet wurde. Dieser Wert kann Werte zwischen 0 und 1 annehmen. Der nächste Wert gibt die Anzahl der aufgrund der chemischen Verschiebungen gefundenen Zuordnungsmöglichkeiten an. Dann folgt eine Angabe über die Größe des Signalvolumens im experimentellen Spektrum. Der nächste Wert gibt die Nummer der Wahrscheinlichkeitstabelle an, die für diese Zuordnung benutzt wurde. Diese Angabe spielt in der zukünftigen Version von *KNOWNOE* keine Rolle mehr, da dort der Zugriff auf die Wahrscheinlichkeitstabellen programmtechnisch anders organisiert ist. Zum Schluss wird der räumliche Abstand (Å) angegeben, den die beiden Atome der Zuordnung in der vorläufigen Modellstruktur haben.

5. *Lower probability limit*

Hier gibt der Benutzer die Mindestwahrscheinlichkeit (0-1) an, die eine Zuordnung für ein bestimmtes NOESY-Signals erreichen muss, um in die Zuordnungsliste übernommen zu werden. Theoretisch sinnvoll sind Wahrscheinlichkeitswerte von größer 0,5. Für die Praxis sind Werte zwischen 0,8 und 1 sinnvoll.

6. *Distance limit*

An dieser Stelle gibt der Benutzer den Abstand in nm an, den beide Atome einer gegebenen Zuordnungsmöglichkeit innerhalb der vorhandenen Modellstruktur nicht überschreiten dürfen, um als Zuordnungsmöglichkeit überhaupt in Frage zu kommen. Am Anfang des iterativen Strukturbestimmungsprozesses mit *AUREMOL* kann man davon ausgehen, dass die gegebene Modellstruktur noch relativ ungenau mit der wirklichen Struktur des zu

untersuchenden Proteins übereinstimmt. So kann es z.B. sein, dass in Wirklichkeit zwei räumlich nahe liegende Atome, die ein starkes NOESY-Signal im experimentellen Spektrum erzeugen, in der vorläufigen Modellstruktur wesentlich weiter auseinanderliegen.
Falls nun der Benutzer einen zu kleinen Suchradius eingestellt hat, besteht die Gefahr, dass die *richtige* Zuordnungsmöglichkeit verworfen wird. Deshalb wird der Benutzer am Anfang einen relativ großen Suchradius, wie z.B.10 nm, einstellen und mit fortlaufender Verfeinerung der Modellstruktur diesen schrittweise bis auf 0,5 nm heruntersetzten (s. Kap. 2.5.2).

7. *Assign limit F1-2D/Assign limit F2-2D*
Hier wird ein Toleranzbereich der chemischen Verschiebung (ppm) für die Frequenzdomänen *F1* und *F2* im Falle von 2D-NOESY-NMR-Spektren angegeben. Der gewählte Toleranzbereich sollte mindestens der digitalen Auflösung des jeweils vorliegenden NMR-Spektrums betragen.
Außerdem sollte er die Unterschiede zwischen den chemischen Verschiebungen gleicher Kerne innerhalb der sequentiellen Zuordnung und dem NOESY-Spektrum berücksichtigen.

8. *Use mutual information*
Optional kann das sog. *NetworkAnchoring* zur zusätzlichen Einschränkung von Zuordnungsmöglichkeiten eingesetzt werden.

9. *Mutual information prob limit*
Hier gibt der Benutzer an, wie stark die Ergebnisse des *NetworkAnchoring* gewichtet werden sollen. Auch beim *NetworkAnchoring* werden Wahrscheinlichkeiten für Zuordnungen von NOESY-Signalen berechnet. Diese Informationen können mit den bereits über das *Bayes'sche* Theorem ermittelten Wahrscheinlichkeiten (s. Kap. 2.5.3) für die einzelnen Zuordnungsmöglichkeiten eines NOESY-Signals verknüpft werden. Dabei wird aus dem vom Benutzer eingegebnen Wert für die Mindestwahrscheinlichkeit (0-1) und der über das *NetworkAnchoring* ermittelten Wahrscheinlichkeit einer Zuordnung ein Faktor berechnet. Mit diesen Faktor wird dann die für die jeweilige Zuordnung berechnete *a priori* Wahrscheinlichkeit (s. Kap. 2.5.3) multipliziert.

10. Assign all peaks
Ordnet allen NOESY-Signalen, welche nicht über die chemischen Verschiebungen oder des *Bayes'schen* Theorem eindeutig zugeordnet werden konnten, diejenige Zuordnungs-möglichkeit zu, die innerhalb der Modellstruktur den jeweils kürzesten Abstand hat. Diese

Option wird in der Regel gegen Ende des Strukturbestimmungsprozesses eingesetzt (s. Kap. 2.5.2).

11. *Assignments to masterlist*

An dieser Stelle kann der Benutzer entscheiden, ob die vom Programm *KNOWNOE* erstellten Zuordnungen für die NOESY-Signale in die *Masterliste* übernommen werden sollen. Die *Masterliste* ist eines der wichtigsten Elemente in der Datenstruktur von *AUREMOL*. In dieser Datei sind alle für die weitere Auswertung benötigten Informationen eines NMR-Spektrums zusammengefasst (Abb. 2.10). Die *Masterliste* lässt sich in zwei Abschnitte gliedern:
Im ersten Abschnitt stehen allgemeine Informationen über das jeweils vorliegende NMR-Spektrum. Dies währen z.B. die Art des Experiments, Länge der Mischzeit während der Aufnahme, spektrale Breite (angegeben in Hz und ppm) der verschiedenen Dimensionen und Angaben über Prozessierungsparameter wie z.B. verwendete Filterfunktionen. Im zweiten Abschnitt sind alle automatisch und vom Benutzer ausgesuchte Signale des NMR-Spektrums aufgeführt. Jedes Signal wird durch folgende Daten charakterisiert:

1. Signalzuordnung mit Angabe des Atomnamens und Sequenzposition.
2. Die chemische Verschiebung bzw. Position (Datenpunkt) des Signals in jeder Dimension innerhalb des Spektrums.
3. Signalintensität.
4. Signalvolumen.
5. Angabe eines Qualitätswerts, der die Wahrscheinlichkeit anzeigt, dass es sich beim betreffendem Signal um ein wirkliches NMR-Signal, Rauschsignal oder Artefakt handelt.

12. *Keep assignments*

An dieser Stelle kann der Benutzer entscheiden, was mit bereits vorhandenen Zuordnungen vor dem Start von *KNOWNOE* geschehen soll. Es gibt drei Optionen:

1. Alle vorhandenen Zuordnungen behalten.
2. Nur manuell erstellte Zuordnungen beibehalten.
3. Alle vorhandenen Zuordnungen entfernen

```
Masterlist (Volumes & intensities rescaled with NC_Proc, scalingfactor sf=2^NC_Proc)
================================================================================

HEADER
        EXPERIMENT:          NOESY->H/H
        DIMENSION:           2
        MIXINGTIME_SEC:      0.03
        AQUISITIONDELAY:     *
        RESOLUTION_W1:       1024
        RESOLUTION_W2:       1024
        RESONANCEFREQUENCY_1:  800.130005
        RESONANCEFREQUENCY_2:  800.130005
        SPECTRAL_WIDTH_HZ_1:   9257.344123
        SPECTRAL_WIDTH_HZ_2:   9257.344123
        SPECTRAL_WIDTH_PPM_1:  11.569800
        SPECTRAL_WIDTH_PPM_2:  11.569800
        OFFSET_PPM_1:        10.933900
        OFFSET_PPM_2:        10.933900
        FILTER_1:            gaussian
        FILTER_2:            gaussian
        LINEBROAD_HZ_1:      1.400
        LINEBROAD_HZ_2:      1.400
        AQ_MODE_1:           simultaneous
        AQ_MODE_2:           simultaneous
        SEQUENCE:
CSP_0_MRGKVKWFDSKKGYGFITKDEGGDVFVHWSAIEMEGFKTLKEGQVVEFEIQEGKKGPQAAHVKVVE
END_HEADER

Total amount of peaks: 3417
Including subpeaks: 3417

PEAKLABEL: HD1 17/HG2 17
COMMENT:
FEATURES: AMBIGUITY 1/1
COMPOUNDS: 1 1
PEAKDESCRIPTION:
      s1       ppm1          s2       ppm2   intensity         volume       prob     width1     width2
     978    -0.11616         953     0.16630       89135         406657     -1.000      0.713      0.740
SUBPEAKS: 0
   -116 166 1000
VOLERR: -1.00 %

...
```

Abbildung 2.10: Ausschnitt aus einer *Masterliste*

3.Material und Methoden

3.1 Software

3.1.1 Benutzte Funktionen im *AUREMOL*

Folgende Funktionen des Programms wurden im Rahmen der Arbeit hauptsächlich benutzt:

1. Ausführung des Programms *KNOWNOE* so wie die Übergabe zu seiner Ausführung benötigter Informationen. Dies wurde über ein entsprechend vorhandenes Eingabefenster der Arbeitsoberfläche von *AUREMOL* bewerkstelligt.

2. Bearbeitung und Darstellung von den 2D-NOESY-NMR-Spektren. Dazu gehört z.B. die Entfernung von nicht interessierenden Signalen oder Spektrenbereichen so wie die Entfernung von bereits in der Zuordnungsliste vorhandenen Zuordnungen vor der Ausführung des Programms *KNOWNOE.*

3. Erzeugung von simulierten 2D-NOESY-NMR-Spektren. Hierbei wurde das Programm-modul *RELAX* benutzt.

3.1.2 Compiler und Programmiersprache

Alle im Rahmen der Arbeit vorgestellten Programme mit dem Suffix „.c“ am Ende des Programmnamens sind von mir in der Programmiersprache *C (ANSI-C)* erstellt worden. Die Programme wurden unter Anwendung der integrierten Entwicklungsumgebung *Visual-Studio 6.0* von der Firma *Microsoft* erstellt. Sowohl die Entwicklung der Programme, wie auch ihre Durchführung wurde entweder an einem PC (*Dell Optiplex GX 260*) oder an einem Notebook (Modell: *Aspire 1405 LC /* Firma: *Acer*), unter dem Betriebssystem *Windows XP (*Firma*: Microsoft),* durchgeführt.

3.2 Teststrukturen

Zu Testzwecken wurden in der Arbeit rückgerechnete 2D-NOESY-NMR-Spektren von den Strukturen der Proteine *Tm*CSP [39] und HPr [40] benutzt. Beide Strukturen wurden mit der Hilfe der NMR-Spektroskopie bestimmt. Im folgendem soll auf die Funktion und Struktur dieser beiden Proteine kurz eingegangen werden.

3.2.1 *Tm*CSP

Beim *Tm*CSP handelt es sich um ein Kälteschockprotein (CSP, → cold-shock protein) aus dem hyperthermophilen Bakterium *Thermotoga maritima*. Kälteschockproteine gehören zu einer Untergruppe von kälteinduzierten Proteinen, welche bei Herabsetzung der Wachstumstemperatur vom betroffenen Organismus exprimiert werden und eine hohe Affinität zu einzelsträngigen Nukleinsäuren besitzen [39]. Bis jetzt ist ihre genaue biologische Funktion jedoch noch unbekannt.

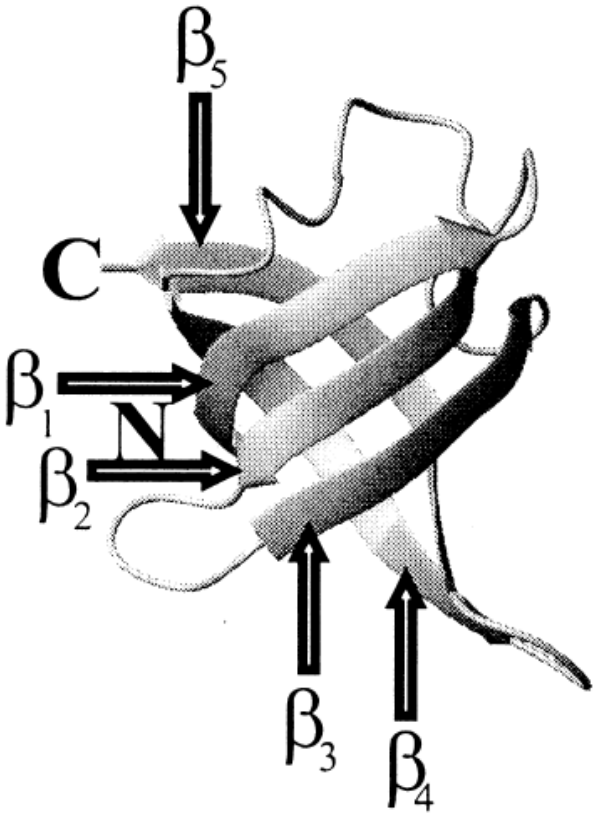


Abbildung 3.1: Bändermodell vom Protein *Tm*CSP

Die Reaktion auf einen Kälteschock mit signifikant niedrigeren Temperaturen als unter physiologischen Bedingungen wirkt sich bei Mikroorganismen auf die Zellwachstumsrate, den Sättigungsgrad an Fettsäuren, sowie die Synthese von DNA, RNA und Proteinen aus

[61]. Zu den bisher strukturell aufgeklärten Kälteschockproteinen gehören, neben *Tm*CSP, CSP*A* aus *Escherichia coli* [62] [63], CSP*B* aus *Bacillus subtilis* und CSP*B* [64] [65] aus *Bacillus caldolyticus* [66]. Diese gehören alle zur sog. *OB*(*oligonucleotide / oligosaccharide binding*) – Faltungsfamilie [67] zu der auch unterschiedliche oligonukleotidbindende und oligosaccharidbindende Proteine gehören. Die β-Faltblätter aller bisher strukturell bekannten CSP*'s* sind in der Form eines griechischen Schlüssels angeordnet und bilden zusammen eine β-Fassstruktur (Abb. 3.1). Die Struktur von *Tm*CSP enthält fünf β-Stränge, die jeweils zwei antiparallele β-Faltblätter bilden. *Tm*CSP ist mit einer Sequenzlängen von 66 Aminosäuren und einer molekularen Masse von 7,474 kDa ein relativ kleines globuläres Protein.

3.2.2 HPr

Beim HPr *(Histidin Containing Phosphocarrier Protein)*, hier aus *Staphylococcus carnosus* [77], handelt es sich, wie der Name schon verrät, um ein histidinhaltiges Phosphotransportprotein. Es setzt sich aus 88 Aminosäuren zusammen. Als wichtigste Strukturelemente sind ein antiparalleles β-Faltblatt, bestehend aus vier Strängen β1,β2, β3, und β4, angeordnet mit der Topologie β1-β4-β2-β3 und drei rechtshändige α-Helices, welche über dem β -Faltblatt platziert sind, zu nennen (Abb. 3.2).

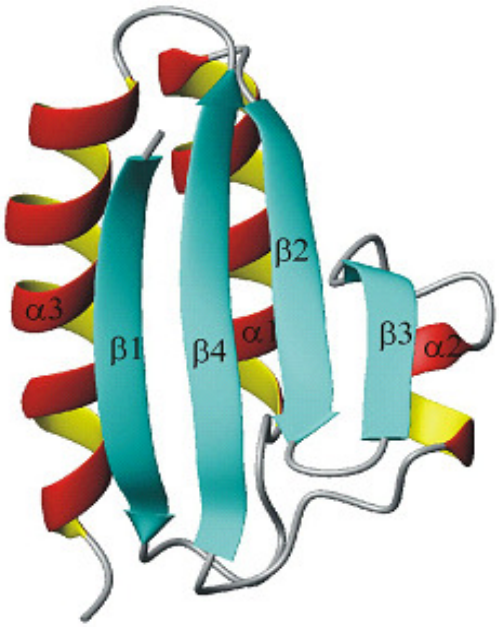


Abbildung 3.2: Bändermodell vom Protein *HPr*

Bei HPr's handelt es sich um kleine monomere Proteine mit einer Molekülmasse von etwa 9,6 kDa. HPr*'s* sind unter anderen aus *Escherichia coli* [69], *Staphylococcus aureus* [70], *Mycoplasma* [71], *Lactobacillus lactis* [72], *Bacillus subtilis* [73], *Salmonella typhimurium*

[74], *Enterococcus faecalis* [75], *Lactobacillus brevis* [76] und *Staphylococcus carnosus* [77] isoliert und charakterisiert worden. Sie spielen eine wichtige Rolle bei der Aufnahme und Posphorylierung von Kohlenhydraten über das *phosphoenolpyrovatabhängige Phosphotransferasesystem* (PTS), und damit bei der Regulation des Kohlenhydrat-methabolismus von Bakterien [68], [78]. Beim *PTS* handelt es sich um ein System zum aktiven Transport von Substraten insbesondere von Kohlenhydraten durch die Zellmembran. HPr*'s* haben hierbei die Aufgabe eine Phosphorylgruppe vom *Enzym I* auf das *Enzym II* zu übertragen (Abb. 3.2). Dabei wird die Phosphorylgruppe kurzeitig an das Stickstoffatom des Histidinrings an der Sequenzposition 15 im aktiven Zentrum des Proteins gebunden.

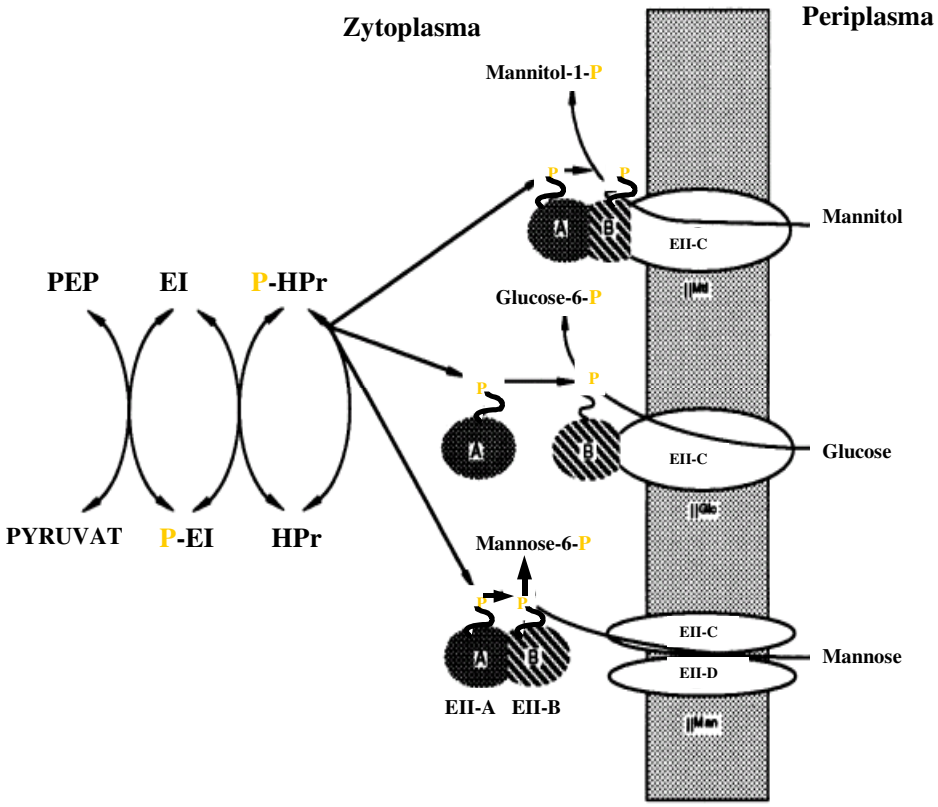


Abbildung 3.3: Phosphoenolpyruvatabhängiges-Phosphotransferasesystem (PTS). Die Grafik [nach 78] zeigt die bei der Phosphorylierung bzw. Transport von Kohlenhydraten beteiligten Proteine E1, EII und *HPr*. Für unterschiedliche Kohlenhydrate gibt es entsprechend verschiedene EIIs. EII setzt sich jeweils aus zwei hydrophilen Domänen, EII-A und EII-B mit jeweils einer Phosphorylierungsstelle, und einer hydrophoben membran-gebundenen Domäne EII-C zusammen. EII-C kann auch in zwei weitere Domänen EII-C und EII-D aufgespalten sein.

3.3 Testspektren

3.3.1 Simulation von 2D-NOESY-NMR-Spektren

Zum Testen des Programms *KNOWNOE* wurden simulierte 2D-NOESY-NMR-Spektren der bereits vorgestellten Proteine *Tm*CSP und HPr erzeugt. Der Grund dafür liegt darin, dass bei simulierten NMR-Spektren die Zuordnung jedes Signals bekannt ist, und somit erst eine Qualitätsbeurteilung einer erstellten Zuordnung bzw. angewandten Zuordnungsmethode möglich ist. Zur Erzeugung der simulierten 2D-NOESY-NMR-Spektren wurde das Programmmodul *RELAX* benutzt. Vor dem Start des Programms wurden folgende Parameter im Eingabedialog, jeweils für beide Proteine, gesetzt:

1.*Cutoff distance* : 0,5 nm
Hierbei handelt es sich um den maximalen Abstand zweier Atome in der Struktur, für die ein NOESY-Signal generiert werden soll. 5 nm beträgt etwa der Maximalabstand zweier Atome, um in einem experimentellen NOESY-NMR-Spektrum ein klar identifizierbares NOESY-Signal zu erzeugen.

2. *Mixing Time*: 0,03 s
Die Mischzeit bezeichnet eine bestimmte Phase während eines 2D-NOESY-NMR-Experimentes, in der Kohärenzen miteinander gekoppelt und in messbare Transversal-magnetisierung umgewandelt werden [29]. Die Mischzeit von 0,03 Sekunden wurde hier relativ kurz gewählt, um den Einfluss der Spindiffusion zu minimieren.

3 *Relaxation Delay:* 0
Ist die Zeit zwischen den einzelnen Pulssequenzen.

4. *Lamor Frequency:* 800,13MHz
Hier wird die Protonenresonanzfrequenz des, während der Rückrechnung theoretisch „benutzten“ NMR-Spektrometer, angegeben.

5. *Sample points in w1* und *Sample points in w2:* 1024
An der Stelle wird die Anzahl der prozessierten Datenpunkte innerhalb der jeweiligen Frequenzdomänen angegeben.

6. *Type of experiment :* NOESY->H/H
Hier gibt der Benutzer die Art des zu simulierenden NMR-Experiments an.

7. *Lineshape: gaussian*
Hier kann der Benutzer die Linienform der zu erzeugenden NOESY-Signale angeben. In diesem Falle wurde die Form einer Gaußlinie gewählt.

8. *Line broadening (H, indirect) /Line broadening (H, direct):* 1.4 Hz
Mit Angabe dieses Parameters lassen sich die Linienbreiten der künstlich erzeugten NOESY-Signale zusätzlich vergrößern, um eine größere Ähnlichkeit mit experimentellen NOESY-Signalen zu erreichen.

9. *Calculate J-Splitting:* no
Es besteht die Möglichkeit Signalaufspaltungen, welche aufgrund von J-Kopplungen entstehen, berechnen zu lassen. Für die in dieser Arbeit zu bearbeitende Fragestellung wurde diese Option nicht benötigt. Neben den oben genannten Parametern, muss vor dem Programmstart die Proteinstruktur, für die das simulierte NMR-Spektrum erstellt werden soll, in Form einer PDB-Datei angegeben werden. Weiter werden noch Angaben über eine *Compound*- und *Meta*-Datei benötigt (s. Kap. 2.5.4). Die unter den genannten Eingangsparametern generierten 2D-NOESY-NMR-Spektren enthielten jeweils 7260 Signale für das Protein *Tm*CSP bzw. 8425 Signale für das Protein HPr.

3.3.2 Nachbearbeitung der Testspektren

Da es bei experimentellen NMR-Spektren oft zu Überlappung von Signalen mit ähnlichen chemischen Verschiebungen kommt, wurden in den simulierten Spektren alle NOESY-Signale zu einem Signal zusammengefasst (aufsummiert), welche sich in beiden Frequenzdomänen um weniger als 0,015 ppm voneinander unterschieden. Dies wurde durch ein selbstgeschriebenes C-Programm (*Summarize_Masterlist.c)* durchgeführt. Das jeweils resultierende NOESY-Signal erhält dabei den Zuordnungsnamen und die chemischen Verschiebungen des volumenmäßig dominierenden Signals. Ziel dieses Vorgehen war realistischere Testbedingungen für das Programm *KNOWNOE* zu schaffen. Nach diesem Schritt reduzierte sich die Gesamtanzahl der NOESY-Signale innerhalb der simulierten Spektren von 7260 auf 5645 (*Tm*CSP) bzw. von 8425 auf 6581 (HPr). Zusätzlich wurden die sich unter der Diagonalen befindlichen bzw. doppelt vorkommenden Signale entfernt, da

diese für die weiteren Auswertungen nicht benötigt wurden (Abb. 3.4). Dies ist mit einer dafür bereits vorhandenen Funktion des Softwarepakets *AUREMOL* durchgeführt worden. Danach enthielten die simulierten NOESY-NMR-Spektren jeweils noch 2736 Signale (*Tm*CSP) bzw. 3178 Signale (HPr).

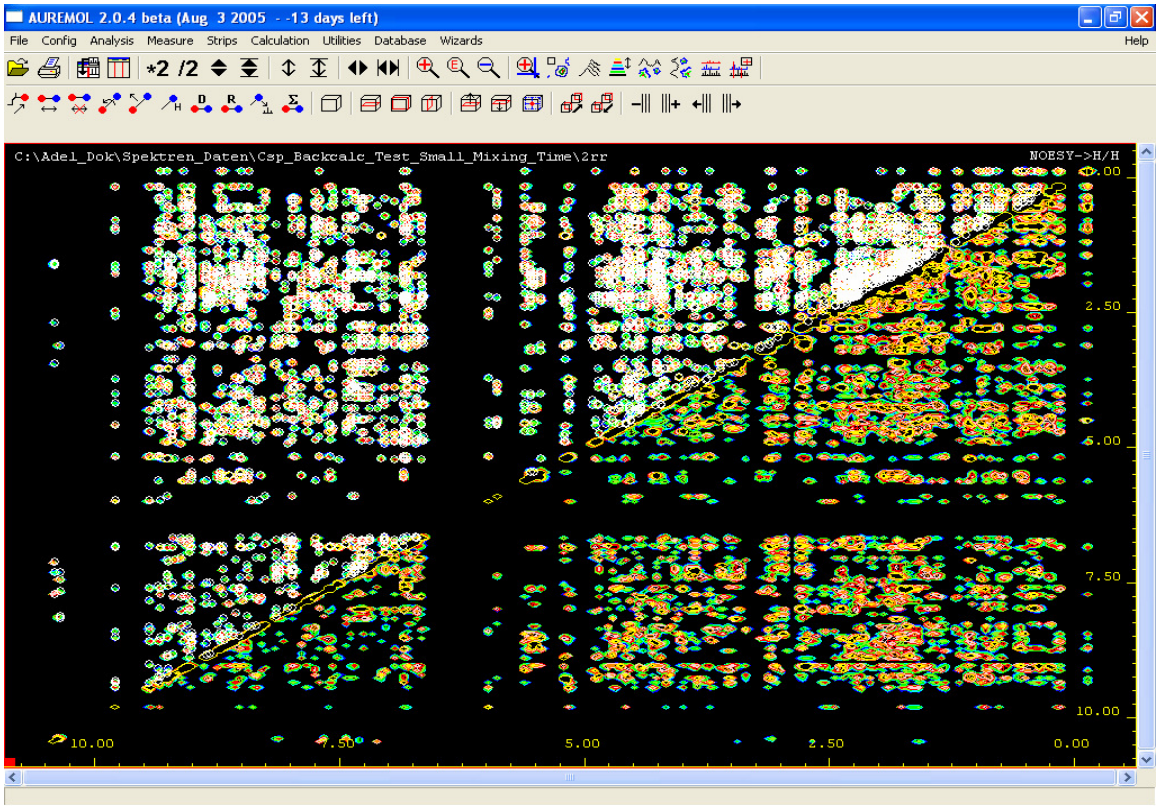


Abbildung 3.4: Darstellung vom simulierten 2D-NOESY-Spektrum des Proteins *Tm*CSP. Die Signale, welche vom Programm *KNOWNOE* bearbeitet werden sollen (obere Hälfte der Diagonalen) sind mit einem weißen Ring markiert.

In Abbildung 3.5 sind die erhaltenen Signale jeweils aufgeschlüsselt nach intraresidualen, sequentiellen, mittelreichweitigen und langreichweitigen NOESY-Signalen dargestellt [13]. Bei *intraresidualen* NOESY-Signalen befinden sich die betreffenden Atome innerhalb der gleichen Aminosäure. Bei *sequentiellen* NOESY-Signalen befinden sich die Atome jeweils in benachbarten Aminosäuren. Bei mittelreichweitigen NOESY-Signalen haben die Atome einen sequentiellen Abstand zwischen zwei- und fünf bzw. bei langreichweitigen NOESY-Signalen mehr als fünf Aminosäuren Abstand voneinander.

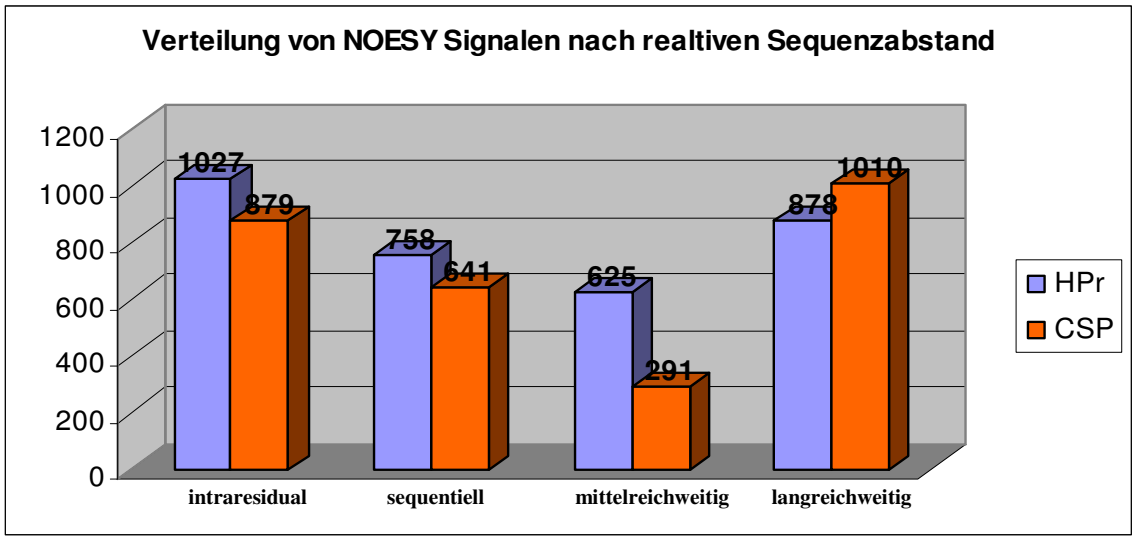


Abbildung 3.5: Anzahl aller NOESY-Signale. Die Grafik zeigt jeweils die Anzahl aller vorhandenen NOESY-Signale der simulierten 2D-NOESY-NMR Spektren von den Proteinen HPr (blau) und *Tm*CSP (orange). Die Signale sind jeweils nach dem relativen Sequenzabstand des volumenmäßig dominanten Atompaares aufgeführt

Aufgrund der hohen Anzahl von Faltblättern, sind beim Protein *Tm*CSP verhältnismäßig viele langreichweitige NOESY-Signale zu beobachten. Beim Protein HPr kann man, wegen der hohen Anteile an helikalen Strukturen, relativ viele mittelreichweitige NOESY-Signale verzeichnen.

3.4 Bekannte Proteinstrukturen als Datenbasis interatomarer Abstände

Als Datenquelle für interatomare Abstände zur Erzeugung umfangreicher Datenbanken aus Wahrscheinlichkeitsdichteverteilungen wurde ein Satz aus 1107 nicht redundanter Proteinstrukturen (PDB-Dateien) [42] benutzt. Alle Strukturen besitzen untereinander eine paarweise Sequenzidentität von weniger als 25%. Dies garantiert einen möglichst breiten Querschnitt der Datenbasis durch alle vorkommenden Typen von Proteinen. 970 der Proteinstrukturen wurden über der Röntgenkristallographie, die restlichen 137 Strukturen mit Hilfe der NMR-Spektroskopie bestimmt. Die durchschnittliche Sequenzlänge der Proteinstrukturen im Datensatz beträgt 271 Aminosäuren.

3.5 Die programmtechnische Erzeugung der neuen Datenbank

Die Erzeugung der neuen Datenbanken aus Wahrscheinlichkeitsdichteverteilungen bestand im Prinzip aus drei wesentlichen Schritten:

1. Extraktion der benötigten Wasserstoffatomkoordinaten aus dem gegebenen Proteinstrukturdatensatz.
2. Ermittlung aller für die Verteilung benötigter Atomabstände.
3. Berechung der Verteilungen aus den ermittelten Atomabständen.

Im folgendem wird auf die genannten Arbeitschritte näher eingegangen.

3.5.1 Extraktion von Wasserstoffatomkoordinaten aus Proteinstrukturen (PDB-Dateien)

Als Datenbasis für die hier erzeugte Datenbank aus Wahrscheinlichkeitsdichteverteilungen wurde ein Datensatz bestehend aus 1107 Proteinstrukturen benutzt. Die Anzahl der generierten Verteilungen betrug über 200 000. In einer bestimmten Wahrscheinlichkeitsdichteverteilung wurden, in Abhängigkeit von der Abstandsklasse, angefangen von einigen 100, bis zu über 200 000 Atomabstände integriert. Um die relativ großen Datenmengen in einen sinnvollen Zeitrahmen und fehlerfrei zu erzeugen bzw. zu verarbeiten, war ein schneller und einfacher Zugriff auf die benötigten Daten notwendig. Dazu mussten zuerst die benötigten Informationen (Atomkoordinaten) aus dem gegebenen Strukturdatensatz extrahiert und in entsprechend geeigneter Form abgespeichert werden. Der direkte Zugriff auf die PDB-Dateien ist vor allem aus folgenden Gründen nicht sinnvoll:

1. Die benötigten Abstände für die Erzeugung einer bestimmten Wahrscheinlichkeitsdichteverteilung sind auf über 1107 einzelnen PDB-Dateien verteilt. Die Zusammenstellung der Daten wäre dadurch, programmtechnisch gesehen, sehr aufwendig und fehleranfällig.

2. PDB-Dateien haben nicht immer das gleiche Format. So stehen z.B. die Koordinatenangaben für die Atome oft in unterschiedlichen Spalten.

3. Es werden oft unterschiedliche Atomnomenklatursysteme benutzt. Dies würde eine zusätzliche Programmierung von Konvertierungsfunktionen für verschiedene Nomenklatursysteme erfordern.

4. Oft sind nicht alle Protonenkoordinaten angegeben oder fehlen vollständig. Dies ist insbesondere bei Röntgenstrukturdaten der Fall.

5. Innerhalb von PDB-Dateien befinden sich viele für diese Arbeit nicht benötigte Daten. Dies erschwert oder verlangsamt den programmtechnischen Zugriff auf die hier interessierenden Atomkoordinaten.

3.5.1.1 Arbeitschritte der Datenextraktion

Die Extraktion der benötigten Wasserstoffatomkoordinaten erfolgte in mehreren Schritten (s. Übersicht in Abbildung 3.6).

1. Aufteilung des Strukturdatensatzes nach NMR- bzw. Röntgen-PDB-Dateien. Da die Datenaufbereitung der NMR-PDB-Dateien zusätzliche Arbeitsschritte erfordert (s. Punkt 4), wurde der vorhandene Strukturdatensatz auf zwei Dateiordner, einer für Röntgen und der andere für NMR-PDB-Dateien, aufgeteilt. Dies wurde mit dem Programm *Split_Xray_NMR.c* (s. Kap. 3.5.1.2) bewerkstelligt.

2. Erzeugung separater PBD-Dateien aus den Strukturmodellen von NMR-PDB-Dateien. Da das im nächsten Schritt benutzte Programm *reduce* [43] nur jeweils das am Anfang stehende Modell einer NMR-PDB-Datei bearbeitet, wurde für alle, jeweils in einer NMR-PDB-Datei stehenden Modelle, eine separate PDB-Datei erstellt. Dies wurde mit dem Programm *Split_Nmr_Pdb .c* (s. Kap. 3.5.1.2) durchgeführt. Aus den vorhandenen 137 NMR-PDB-Dateien wurden somit insgesamt 2735 neue PDB-Dateien erzeugt.

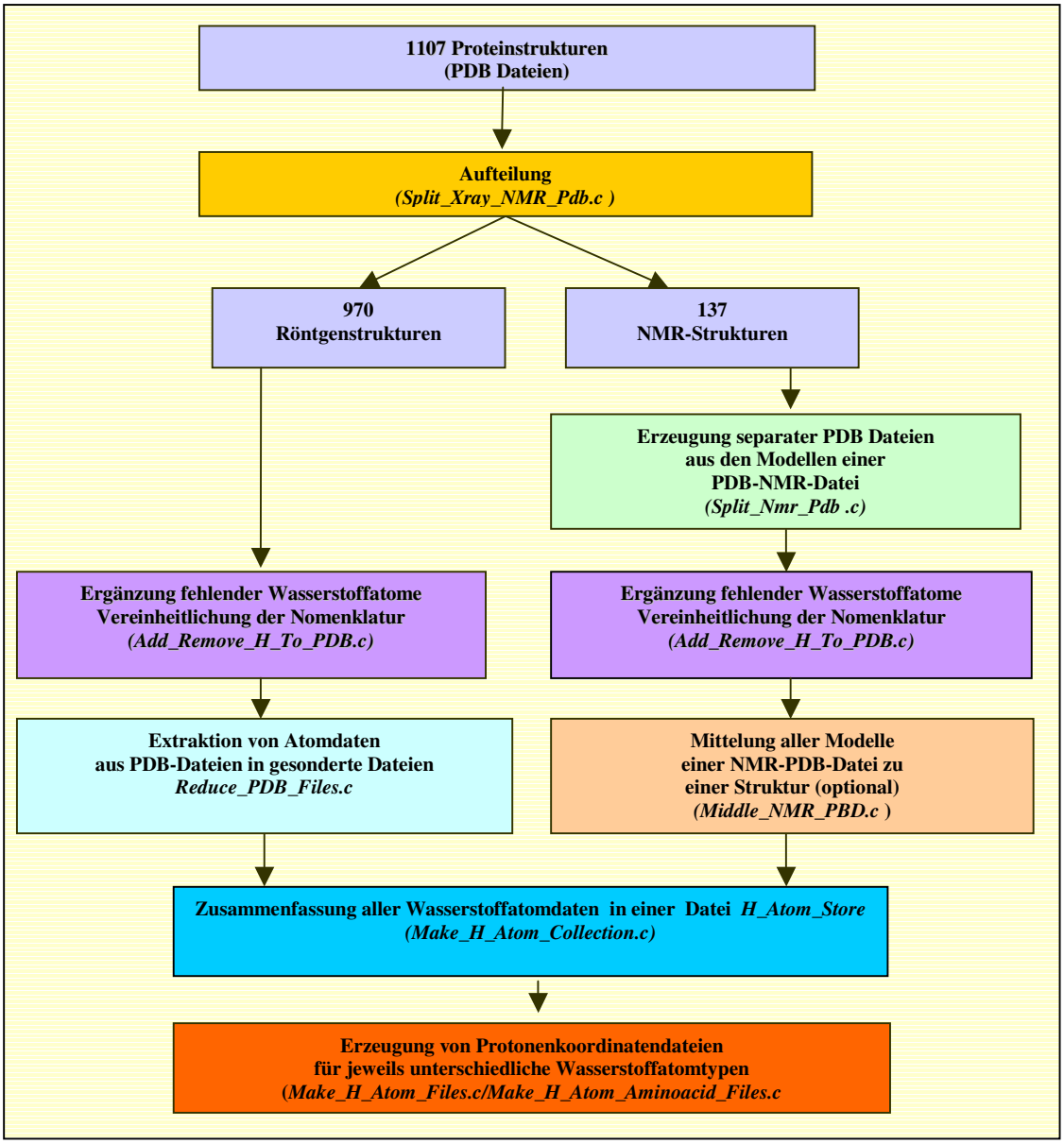


Abbildung 3.6: Extraktion von Wasserstoffatomdaten aus PDB-Dateien. Das für einen bestimmten Arbeitschritt benutzte Programm ist kursiv in Klammern angegeben.

3. Ergänzung fehlender Wasserstoffatome und Vereinheitlichung der Atomnomenklatur. Der nächste Schritt war die Ergänzung fehlender Wasserstoffatomkoordinaten und die Vereinheitlichung der Atomnamennomenklatur. Dafür wurde das Programm *Add_Remove_H_To_PDB.c* benutzt, in welches das Programm *reduce* integriert wurde. Das Programm *reduce* ist in der Lage sowohl innerhalb von Proteinstrukturdaten (PDB-Dateien) wie auch in Nukleinsäurestrukturdaten fehlende Wasserstoffatomkoordinaten zu ergänzen bzw. bereits vorhandene Wasserstoffatome zu entfernen.

4. Geometrische Mittelung der NMR-Strukturen. Für jede bestimmte Abstandsklasse sind jeweils zwei Wahrscheinlichkeitsdichteverteilungen erzeugt worden. Der Unterschied zwischen den beiden Verteilungen besteht in der Art und Weise in der die Abstandsdaten aus den gegebenen NMR–Strukturen jeweils integriert wurden. Im ersten Fall wurden nur die Atomabstände aus dem ersten in der jeweiligen NMR-PDB-Datei aufgeführten Strukturmodelle benutzt. Um allerdings keine strukturellen Informationen zu verlieren, wurden im zweiten Fall alle Strukturmodelle aus den einzelnen NMR-Strukturen zuvor zu einer einzigen Struktur durch Mittelung der Raumkoordinaten der entsprechenden Atome zusammengefasst. Die geometrische Mittelung wurde mit dem Programm *Middle_NMR_PBD.c* (s. Kap. 3.5.1.2) durchgeführt. Bei dieser Vorgehensweise besteht allerdings die Gefahr der Entstehung unrealistischer interatomarer Atomabstände innerhalb der resultierenden Struktur. Aufgrund dessen kann der zweite Fall nicht als optimale Lösung angesehen werden, sondern nur als optionale Alternative.

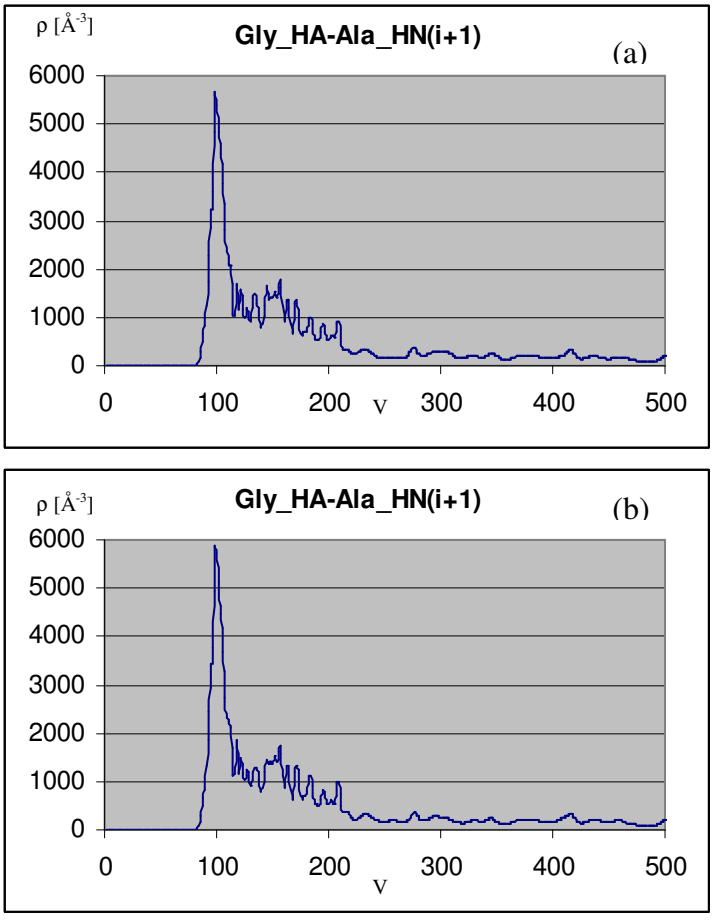


Abbildung 3.7: Auswirkung geometrischer Mittelung. Die Abbildungen zeigen jeweils die Volumenwahrscheinlichkeitsdichteverteilungen für die Abstandsklasse [HA, Gly / HN, Ala (i+0)]. Verteilung *b* unterscheidet sich von der Verteilung a nur darin, dass bei ihrer Erzeugung die Modelle der vorhandenen NMR-Strukturen, vor der Bestimmung der Atomabstände, zu einer Struktur geometrisch gemittelt wurden.

Es hat sich allerdings gezeigt, dass die zwei genannten Vorgehensweisen zur Einbeziehung von NMR-Strukturen in die Wahrscheinlichkeitsdichteverteilungen praktisch keinen Einfluss auf den resultierenden Kurvenverlauf haben (Abb. 3.7). Die Gründe hierfür liegen zu einem darin, dass der Anteil der vorhandenen NMR-Strukturen am gesamten Strukturdatensatz nur etwa 12% ausmacht, und zum anderen, dass das hier benutzte Kurvenglättungsverfahren sog. *statistische Ausreißer* weitgehend unterdrückt. Für die hier durchgeführten Testreihen wurden jeweils die Wahrscheinlichkeitsverteilungen mit den geometrisch gemittelten NMR-Strukturen benutzt. Eine bessere Vorgehensweise zur Integration mehrere Modelle einer NMR-Struktur würde darin bestehen, alle für ein gesuchtes Atompaar innerhalb der Strukturmodelle gefundenen Abstände zuerst zu bestimmen und diese zu einem Abstand zu mitteln. Das Verfahren wurde allerdings aus verschiedenen Gründen nicht angewandt. Zu einem ist das genannte Verfahren aufgrund der vorhandenen Programmstrukturen programmtechnisch wesentlich aufwendiger zu realisieren. Außerdem würde die Rechenzeit zur Generierung einer Verteilung stark ansteigen, da sich hierbei die Anzahl der nach bestimmten Abständen zu durchsuchenden Strukturen mehr als verdreifachen (über 3700 anstatt 1107) würde. Weiter ist, aus den bereits weiter oben genannten Gründen, keine signifikante Änderung der Kurvenverläufe zu erwarten.

5. Extraktion von Atomkoordinaten aus Röntgen-PDB-Dateien. Im nächsten Arbeitschritt wird für jede Röntgen-PDB-Datei eine gesonderte Datei erzeugt, welche nur noch die für die weiteren Schritte benötigten Informationen der entsprechenden Proteinstruktur enthält. Dazu zählen die Atomnamen, die Namen der zugehörigen Aminosäuren, die jeweilige Sequenzposition sowie die zugehörigen Raumkoordinaten. Abbildung 3.8 zeigt einen Ausschnitt aus dem Inhalt einer solchen Datei. Zur Ausführung des hier genannten Arbeitschrittes wurde das Programm *Reduce_PDB_Files.c* geschrieben.

```
...
..
1HZ  LYS   108    20.870   30.224   4.804
2HZ  LYS   108    21.072   29.304   3.470
3HZ  LYS   108    21.766   30.780   3.557
N    MET   109    18.611   30.445   -3.503
CA   MET   109    18.406   30.017   -4.854
C    MET   109    17.300   30.816   -5.497
```

Abbildung 3.8: Extrahierte Atomdaten einer PDB-Datei. In den Spalten, von links nach rechts, stehen jeweils der Atomname, Aminosäurename, Sequenzposition und die entsprechenden Raumkoordinaten x, y und z.

6. Zusammenfassung aller extrahierten Atomkoordinaten. Hier wurde eine Datei erzeugt *(H_Atom_Store*) in der alle Daten aus dem in Arbeitschritten 4 und 5 erzeugten Dateien zusammengefasst wurden. Abbildung 3.8 zeigt einen Ausschnitt aus der Datei *H_Atom_Store*.

```
Structure 1
          HA  VAL   371    1.766    -3.346   342.404
HB  VAL   371    3.720    -1.669   342.264
1HG1 VAL   371    2.813    -0.302   344.148
2HG1 VAL   371    1.498    -1.119   343.236
3HG1 VAL   371    2.027    -1.768   344.825
1HG2 VAL   371    5.068    -1.353   344.333
2HG2 VAL   371    4.338    -2.837   345.034
3HG2 VAL   371    5.383    -2.952   343.577
...
Structure 2
HA  LYS   2      8.057    30.823   18.835
1HB  LYS   2      6.752    31.573   20.505
2HB  LYS   2      7.923    32.821   20.953
..
Structure 1107
```

Abbildung 3.8: Ausschnitt aus der Datei H_Atom_Store

7. Erzeugung von Protonenkoordinatendateien für unterschiedliche Wasserstoffatome. Im letzten Schritt der Datenextraktion wurde für verschiedene in Proteinen vorkommende Wasserstoffatome jeweils eine gesonderte Datei (*Protonenkoordinatendatei*) für die jeweils zugehörigen räumlichen Koordinaten angelegt. Dies hat den Vorteil, dass man bei der Suche nach bestimmten Atomabständen, anstatt des gesamten Strukturdatensatz, nur die Protonenkoordinatendateien der interessierenden Wasserstoffatome durchsuchen muss. Für die Durchführung dieses Arbeitschrittes wurden die Programme *Make_H_Atom_Files.c* und *Make_H_Atom_Aminoacid_Files.c* geschrieben. Abbildung 3.9 zeigt einen Ausschnitt aus einer *Protonenkoordinatendate*i. Wie man sieht, sind die Protonenkoordinaten in hintereinanderliegenden Blöcken abgespeichert. Jeder Block enthält die Sequenzposition und die Raumkoordinaten (*x, y, z*) eines Wasserstoffatomtyps aus einer bestimmten Struktur und ist jeweils mit einer laufenden Nummer versehen. Der Dateiname einer bestimmten Protonenkoordinatendatei setzt sich dabei entweder nur aus dem Wasserstoffatomnamen (nach IUPAC) oder noch zusätzlich aus dem Namen der zugehörigen Aminosäure (Dreibuchstabencode) zusammen (z.B. *Arg_HA*). Es wurden somit zwei Arten von Protonenkoordinatendateien erzeugt:

```
Structure 1
377 2.294 -6.795 332.335
390 10.577 -3.471 316.318
398 8.995 0.906 307.835
424 3.048 7.937 287.684
430 -3.902 6.202 280.067
439 -1.428 4.916 265.267
452 -8.938 1.173 246.940
456 -8.311 0.532 241.413
462 3.952 -6.849 229.798
468 6.614 -3.359 221.980
Structure 2
4 11.445 32.732 13.281
7 6.636 33.872 10.652
17 -2.901 29.547 1.165
36 0.842 20.914 2.818
39 1.558 21.592 -2.394
59 -0.527 10.686 4.339
66 7.314 15.046 -1.792
70 12.664 17.264 -2.905
Structure 3
12 81.367 28.065 33.622
29 69.785 14.806 32.146
31 72.096 18.116 37.666
..
```

Abbildung 3.9: Ausschnitt aus einer *Protonenkoordinatendatei* (hier für die *HA*-Atome der Aminosäure Alanin)

Die erste Art enthält jeweils die Raumkoordinaten für einen bestimmten Wasserstoffatomtyp (z.B. *HA, HB*.) ohne seine Zugehörigkeit zu einer bestimmten Aminosäure zu berücksichtigen. Es wurden soviel *Protonenkoordinatendateien* dieser Art erzeugt, wie es nach IUPAC unterschiedliche Wasserstoffatomnamen in den 20 natürlichen Aminosäuren gibt (43) (Anhang *C*). Bei der zweiten Art von *Protonenkoordinatendatei* wurde zusätzlich noch die Aminsäurezugehörigkeit der Protonen berücksichtigt. Dadurch lassen sich 160 verschiedene Wasserstoffatomtypen (s. Anhang) definieren (z.B. *HA* in Glycin, *HA* in Arginin, *HB3* in Serin usw.) Insgesamt wurden 156 dieser Dateien erzeugt. Für die Wasserstoffatome *HE2* in Histidin, *HD2* in Aspartat, *HE2* in Glutamat und *HD1* in Histidin konnten keine *Protonenkoordinatendateien* erzeugt werden. Der Grund dafür ist, dass das Programm *reduce* diese Protonen zu den betreffenden Aminosäuren nicht hinzuaddiert, da diese unter physiologischen Bedingungen (pH 7) deprotoniert vorliegen.

3.5.1.2 Programme zur Datenextraktion

Im folgendem wird auf die wesentliche Funktionsweise der benötigten Programme zur Datenextraktion in der Reihenfolge ihrer Ausführung kurz eingegangen. Alle Programme wurden von mir in der Programmiersprache C (*ANSI C*) geschrieben. Eine Ausnahme ist das Programm *reduce* [43]. Es wurde über das Internet bezogen und ist in der Programmiersprache C++ verfasst worden.

1. *Split_Xray_NMR_Pdb.c*

Das Programm ist in der Lage Röntgenstrukturdaten von NMR-Strukturdaten zu unterscheiden. Dabei sucht das Programm nach bestimmten Schlüsselwörtern wie z.B. „*resolution*“ oder „*modell*“ innerhalb der in Frage stehenden PDB-Datei.

2. *Add_Remove_H_To_PDB.c*

Das Programm addiert oder entfernt Wasserstoffatome in standardisierter Geometrie zu bzw. von PDB-Dateien. Dabei übergibt der Benutzer dem Programm eine Namensliste der zu bearbeitenden PDB-Dateien. Das Programm läd jeweils eine Datei in den Speicher und übergibt sie an das Programm *reduce* zur Bearbeitung weiter. Danach liest das Programm die nächste PDB-Datei ein. Auf diese Weise ist es möglich viele PDB-Dateien automatisch mit dem Programm *reduce* bearbeiten zu lassen.

3. *Split_NMR_PDB.c*

Legt für jedes in einer NMR-PDB-Datei vorhandene Strukturmodell eine separate PDB-Datei an. Der Name einer solchen Datei setzt sich dabei aus dem Namen der ursprünglichen PDB-Datei und der laufenden Nummer des betreffenden Strukturmodells zusammen (z.B. *pdb1aab.ent_modell_3*).

4. *Middle_NMR_PBD.c*

Erzeugt eine geometrisch gemittelte Struktur aus allen in einer bestimmten NMR-PDB-Datei vorhandenen Modellstrukturen. Die so erzeugte Struktur wird in einer neuen Datei abgespeichert und erhält dabei den Namen der ursprünglichen NMR-PDB-Datei mit dem Suffix „_ *middled* „ (z.B. *pdb1aab.ent_middled*). Bei dieser Art von Dateien sind bereits für die weiteren Schritte nicht benötigte Daten wie der Dateikopf oder Kommentare von dem Programm entfernt worden.

5. *Reduce_PDB_Files.c*

Dieses Programm extrahiert aus einer oder mehreren PDB-Dateien die für die Generierung von Wahrscheinlichkeitsdichteverteilungen benötigten Atomkoordinaten. Zu den benötigten Daten zählen in diesen Zusammenhang die Atomnamen, ihre Sequenzpositionen, ihre Aminosäurezugehörigkeit und die zugehörigen dreidimensionalen Koordinaten. Der Name der hier erzeugten Datei setzt sich aus dem Namen der ursprünglichen PDB-Datei und dem Suffix „*_reduced*“ zusammen (z.B. *pdb1aa0_ent_reduced).*

6. *Make_H_Atom_Collection.c*

Fasst alle aus den gegebenen Strukturdatensatz benötigten Informationen in einer einzigen Datei (*H_Atom_Store*) zusammen. Als Datengrundlage dienen die mit den Programmen *Reduce_PDB_Files.c* und *Middle_NMR_PBD.c* generierten Dateien.

7. *Make_H_Atom_Files.c*

Fasst alle dreidimensionalen Koordinaten und Sequenzpositionen für jeweils einen bestimmten Wasserstoffatomtyp aus allen 1107 vorhandenen Proteinstrukturen in einer separaten Datei (*Protonenkoordinatendatei*) zusammen. Hierbei wird ein Wasserstoffatom durch seinen Atomnamen (z.B. *HA*) charakterisiert. Als Datenbasis zur Generierung dieser Dateien diente die vom Programm *Make_H_Atom_Collection.c* generierte Datei *H_Atom_Store*. Die erstellten *Protonenkoordinatendateien* dienten letztendlich als Datenbasis für die Erzeugung der hier erzeugten Datenbanken aus Wahrscheinlichkeitsdichte-verteilungen.

8. *Make_H_Atom_Aminoacid_Files.c*

Das Programm hat dieselbe Funktion wie das Programm *Make_H_Atom_Files.c.* Allerdings wird hier ein Wasserstoffatom durch den Atomnamen und zusätzlich noch durch seine Aminosäurezugehörigkeit charakterisiert (z.B. H in Glycin).

3.5.2 Berechnung von Wahrscheinlichkeitsdichte-verteilungen

Für die Generierung der Volumen– und Abstandswahrscheinlichkeitsdichteverteilungen wurde das Programm *Calc_Prob_Tabs.c* geschrieben. Das Programm hat in wesentlichen zwei Funktionen:

1. Die Bestimmung aller zu einer bestimmten Abstandsklasse zugehörigen Atomabstände aus dem gegebenen Strukturdatensatz.
2. Die Erzeugung von Volumen- bzw. Abstandswahrscheinlichkeitsdichteverteilungen aus Atomabständen.

3.5.2.1 Effektive Akquisition von Atomabständen

Für die Bestimmung aller Atomabstände einer bestimmten gegebenen Abstandklasse führt das Programm *Calc_Prob_Tabs.c* in wesentlichen folgende drei Arbeitschritte aus:

1. Öffnen zweier *Protonenkoordinatendateien.*
2. Abspeicherung der Daten (dreidimensionale Koordinaten) einer *Protonenkoordinatendatei* in drei dreidimensionale *Arrays*[2].
3. Durchsuchen des *Arrays* nach den jeweils interessierenden Atompaaren und Berechnung ihrer Abstände.

Als Datenbasis für Atomabstände benutzt das Programm *Calc_Prob_Tabs.c* die bereits beschrieben *Protonenkoordinatendateien*. Diese ermöglichen, wie bereits erwähnt, einen effektiven Datenzugriff. Bei der Bestimmung von Atomabständen einer bestimmten Abstandsklasse benutzt das Programm nur diejenigen zwei Protonenkoordinatendateien, die den Atomen des gefragten Atompaares entsprechen. Dadurch wird das Durchsuchen nicht benötigter Daten vermieden. Die Abspeicherung der Daten einer der beiden benötigten *Protonenkoordinatendateien* innerhalb eines *Arrays* erleichtert die Suche nach bestimmten Atompaaren erheblich. Hierbei wird jede der dreidimensionalen Koordinaten (x, y und z) eines Atoms (das zweite Atom des gesuchten Paares) jeweils in einem dreidimensionalen *Array* abgespeichert (Abb. 3.9). Die Werte wurden folgendermaßen abgespeichert:

[2] Ein *Array* ist die Bezeichnung einer bestimmten Datenstruktur oder Datenfeldes in der Informatik

1. *Dimension*: Gibt an, ob es sich um x, y, oder z-Koordinaten handelt (0=x, 1=y, 2=z)

2. *Dimension*: Nummer der Struktur, in der sich das Atom befindet.

3. *Dimension*: Sequenzposition innerhalb des Proteins

$$[0]\ [49]\ [399] = x_i$$
$$[1]\ [49]\ [399] = y_i$$
$$[2]\ [49]\ [399] = z_i$$

Abbildung 3.9: Abspeicherung von Raumkoordinaten in drei dreidimensionale *Arrays*.

Um nun für die Abstandsberechnung die Koordinaten für ein bestimmtes Atom gesuchten Partner (falls vorhanden) zu finden, muss nur der Inhalt der zugehörigen drei Speicherzellen im Array abgefragt werden. Dies ist deshalb möglich, da die Eigenschaften des jeweils gesuchten Atompartners (Strukturnummer und Sequenzposition) die Adresse seiner dreidimensionalen Koordinaten im Array bestimmen. Falls der Inhalt der Speicherzellen leer ist (Wert = 0), kann man davon ausgehen, dass es keinen passenden Partner für das jeweils in Frage stehende Atom innerhalb der gerade durchsuchten Struktur gibt. Durch dieses Vorgehen wird vermieden, den gesamten Datensatz unnötigerweise zu durchsuchen.

3.5.2.2 Reduzierung großer Wertemengen

Für bestimmte Abstandsklassen bzw. Atompaare wurden, aufgrund der großen Strukturdatenbasis, oft sehr große Wertemengen (n>1000 000) erhalten. Dies führte oft zu sehr langen Rechenzeiten während der Berechnung einer Verteilung. In Anbetracht der großen Anzahl (n>200 000) der im Rahmen der Arbeit zu erzeugender Verteilungen, würde die Generierung der gesamten Datenbank (*Datenbank 1- 3*) unakzeptabel viel Zeit (mehrere Woche) beanspruchen. Es musste deshalb ein Weg gefunden werden, die relativ großen Datenmengen, vor der weiteren Verarbeitung und ohne nennenswerten Informationsverlust oder gar Weglassen von Werten, zu reduzieren. Die Lösung bestand darin, die für eine bestimmte Abstandsklasse gefundenen Werte vorher, entsprechend ihrer Größe, zu bestimmten Abstands- bzw. Volumenintervallen zuzuordnen. Die Teilintervalle befinden sich innerhalb des angegebenen Wertebereiches (Volumenverteilung: 0-0,042 $Å^3$ / Abstandsverteilung: 1,7Å-100Å) einer Wahrscheinlichkeitsdichteverteilung. Für alle gebildeten j Teilintervalle gilt:

$$[a_j,\ b_j\ ] \quad \text{mit } (j=1,2,..J)$$

und

$$a_j < b_j\ ,\ \ a_{j+1} = b_j\ ,\ \ a_{j+1} < b_{j+1}$$

Aus allen gefundenen Werten (Abstände und Volumina) E_i, die sich innerhalb eines bestimmten Intervalls befanden, wurde nun ein Mittelwert $\bar{E}_j$ berechnet. Für die Berechnung einer Wahrscheinlichkeitsdichteverteilung wurde die nun geringere Anzahl von Mittelwerten eingesetzt. Bei der Wahl der Intervallbreiten mussten folgende Aspekte berücksichtigt werden.

1. Die Intervallbreiten mussten so gewählt werden, dass sich die aus den Mittelwerten resultierenden Verteilungen nicht wesentlich von den Verteilungen unterscheiden, welche aus den Originalwerten generiert werden. Das bedeutet, dass beispielsweise keine vorhandenen Minima oder Maxima durch die Mittelwertbildung verloren gehen oder miteinander im Kurvenverlauf verschmelzen durften.

2. Zu groß gewählte Intervalle können zu erheblichen Informationsverlust führen. Dies gilt besonders für Bereiche, in denen sich besonders viele Werte befinden oder sich relativ schnell ändern.

3. Zu klein gewählte Intervallbreiten können in Bereichen mit relativ wenig Werten zu leeren Gruppen bzw. Mittelwerten mit dem Wert Null führen.

Die Punkte 1. und 2. können vor allem Volumenwahrscheinlichkeitsdichteverteilungen betreffen bei denen sich die Werte einer bestimmten Abstandsklasse oft relativ ungleichmäßig auf den gesamten Wertebereich verteilen (s. Kap. 4.1.2.1). Als Orientierung bei der Wahl der Intervallbreiten dienten hierbei Abstandshäufigkeitsverteilungen zwischen *HA*-Atompaaren mit jeweils sequentiellen Abständen von einer, zwei, drei, vier, acht und mehr als acht Aminosäuren (Abb. 3.10). Die Verteilungen geben in etwa einen Überblick darüber, in welchen Bereichen sich Abstände verschiedener Abstandsklassen verteilen. Wie man aus den Abbildungen entnehmen kann, häufen sich im Bereich von etwa 4-5 Å hauptsächlich Abstände von Atomen aus jeweils benachbarten Aminosäuren. Im Bereich von etwa 5-15 Å befinden sich vorwiegend Atomabstände von Atompaaren mit zwei bis acht Aminosäuren Abstand innerhalb der Proteinsequenz. Im Bereich von etwa größer als 15 Å sind meist Abstände von Atompaaren mit jeweils sequentiellen Abständen von mehr als acht Aminosäuren vertreten.

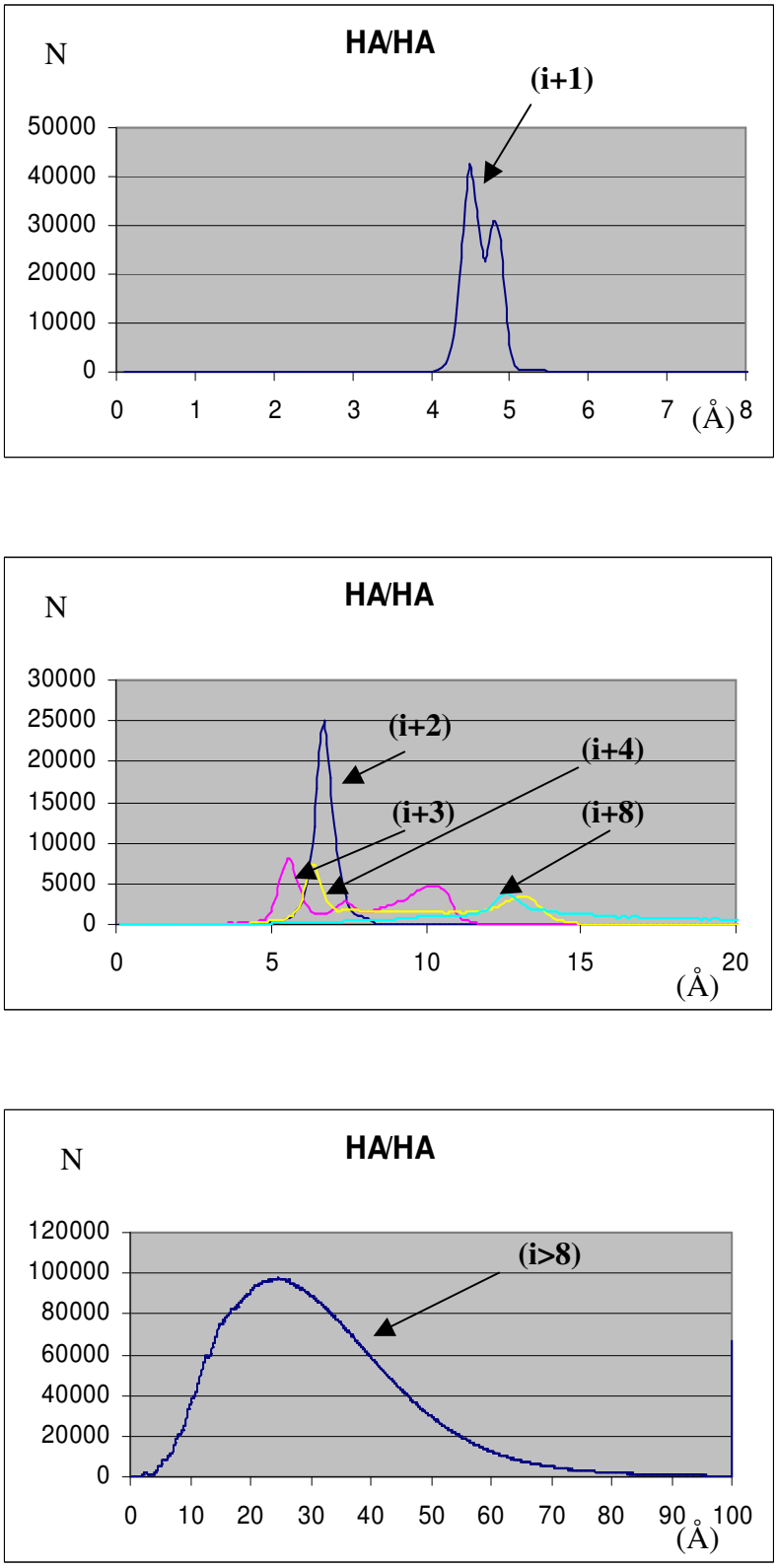


Abbildung 3.10: Abstandshäufigkeitsverteilungen zwischen *HA*-Atompaaren. In den oberen beiden Grafiken sind Abstandshäufigkeitsverteilungen von *HA* Atompaaren für sequentielle Abstände von 1, 2, 3, 4 und 8 Aminosäuren zu sehen. Die untere Grafik zeigt eine Verteilung in der alle Atomabstände von *HA* Atomen mit jeweils mehr als 8 Aminosäuren Abstand innerhalb der Proteinsequenz zusammengefasst wurden. Ein Datenpunkt entspricht der Anzahl der ermittelten Atomabstände innerhalb eines Intervalls der Größe 0,1 Å.

Es hat sich im Hinblick auf diese Verteilungskurven als sinnvoll erwiesen, den gesamten interessierenden Abstandbereich (1,7Å - ∞) in drei Bereiche zu unterteilen:

1. 1,7 Å –4 Å /5 Å (entspricht Volumenbereich: 4,2 x 10^{-2} Å^3 - 4 x 10^{-4} Å^3 / 6,4 x 10^{-5} Å^3)
2. 4 Å /5 Å-30 Å (entspricht Volumenbereich: 4 x 10^{-4} Å^3 / 6,4 x 10^{-5} Å^3 - 1,37 x 10^{-9} Å^3)
3. 30 Å-∞ Å (entspricht Volumenbereich: 1,37 x 10^{-9} Å^3 – 0 Å^3)

Die in Klammern angegebenen Volumenbereiche entsprechen den mit der Formel $V=(1/r^6)$* Å^9 umgerechneten Grenzen der Abstandsintervalle. Im Fall einer Gruppenbildung, wurden die genannten Bereiche in äquidistante Intervalle unterteilt. Dabei wird meist die Anzahl der gebildeten Intervalle in den Bereichen größer bzw. kleiner gewählt, in denen die betreffende Verteilung weniger bzw. mehr Struktur besitzt. Durch diese Vorgehensweise lies sich die für eine bestimmte Verteilung vorhandene Anzahl zu verarbeitender Werte, ohne nennenswerten Informationsverlust, minimieren. Die Anzahl der gebildeten Intervalle hing unter anderen noch von weiteren Faktoren ab wie der Menge der jeweils zu erzeugenden Verteilungen und der durchschnittlich gefundenen Abstände bzw. Volumina der jeweiligen Abstandklassen. In Abbildung 3.12 sind die gewählten Intervallbreiten jeweils für Abstands- und Volumenwahrscheinlichkeitsdichteverteilungen für alle in dieser Arbeit erzeugten Typen von Abstandsklassen aufgeführt. Die Anzahl der Werte die sich innerhalb eines bestimmten Intervalls befinden wird bei der Berechnung der Wahrscheinlichkeitsdichtverteilung (s. nächstes Kapitel) in Form eines Gewichtungsfaktors (g_j) berücksichtigt. Diese Parameter mussten deshalb während der Mittelwertbildung gesondert gespeichert werden. In Abbildung 3.11 ist die beschriebene Vorgehensweise zur Reduzierung großer Wertemengen durch Mittelwertbildung noch einmal beispielhaft verdeutlicht.

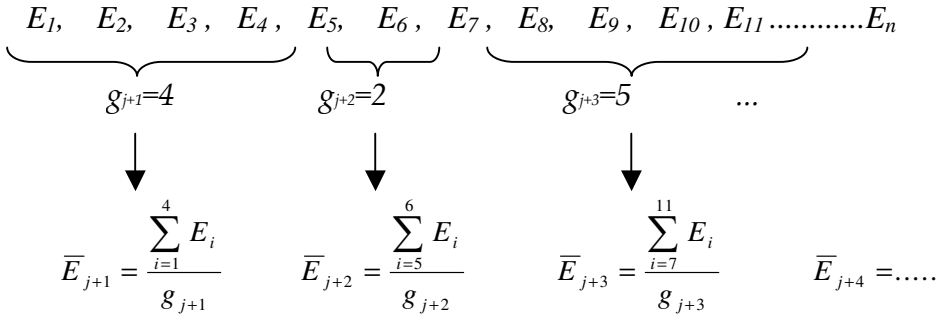


Abbildung 3.11: Reduzierung einer Wertemenge durch Mittelwertbildung. Die Anzahl einer gegebenen Wertemenge *E* der Größe *n* wird durch Zusammenfassen von jeweils hintereinanderliegenden Werten der Anzahl g_j zu Mittelwerten $\bar{E}_j$ verkleinert.

	Volumen-wahrscheinlichkeits-dichteverteilungen		Abstands-wahrscheinlichkeits-dichteverteilungen	
Datenbank	**Wertebereich [Å^3]**	**Anzahl äquidistanter Intervalle**	**Wertebereich [Å]**	**Anzahl äquidistanter Intervalle**
Datenbank 1 ΔS={0,1,2,3...,8}	$0 - 1.37*10^{-9}$ $1.37*10^{-9} – 6.4*10^{-6}$ $6.4*10^{-6} – 4.2*10^{-2}$	500 500 500	1.7Å-100 Å	983 Intervallbreite =0.1Å
Datenbank 1 ΔS>8	$0 - 1.37*10^{-9}$ $1.37*10^{-9} – 6.4*10^{-6}$ $6.4*10^{-6} – 4.2*10^{-2}$	500 500 500	1.7Å-100 Å	983 Intervallbreite =0.1Å
Datenbank 2 ΔS={0,1,2,3...,8}	$0 - 1.37*10^{-9}$ $1.37*10^{-9} –2.44*10^{-4}$ $2.44*10^{-4}– 4.2*10^{-2}$	1000 20 000 Keine	1.7Å-100 Å	983 Intervallbreite =0.1Å
Datenbank 2 ΔS>8	$0 - 1.37*10^{-9}$ $1.37*10^{-9} – 6.4*10^{-6}$ $6.4*10^{-6} – 4.2*10^{-2}$	500 500 500	1.7Å-100 Å	983 Intervallbreite =0.1Å
Datenbank 3 ΔS={0,1,2,3...,8}	$0 - 1.37*10^{-9}$ $1.37*10^{-9} – 6.4*10^{-6}$ $6.4*10^{-6} – 4.2*10^{-2}$	Keine Keine Keine	1.7Å-100 Å	983 Intervallbreite =0.1Å
Datenbank 3 ΔS>8	$0 - 1.37*10^{-9}$ $1.37*10^{-9} – 2.44*10^{-4}$ $2.44*10^{-4}– 4.2*10^{-2}$	100 500 Keine	1.7Å-100 Å	983 Intervallbreite =0.1Å

Abbildung 3.12: Übersicht der Gruppenbildung von Werten (Abstände/Volumina). "Keine" bedeutet in diesen Zusammenhang, dass gefundene Werte innerhalb des entsprechenden Wertebereich nicht in Klassen bzw. zu Gruppen für jeweils bestimmte Teilintervalle des Bereichs zugeordnet bzw. zusammengefasst wurden.

3.5.2.3 Berechnung von Verteilungskurven

Aus den für eine bestimmte Abstandsklasse erhaltenen Werten (Abstände/Signalvolumina) E_i bzw. Mittelwerten $\bar{E}_j$ (s. vorheriges Kapitel) werden nun die entsprechenden Wahrscheinlichkeitsdichteverteilungen durch Aufsummierung von normierten Gaußkurven angenähert [100] werden. Dabei führt das Programm folgende Schritte nacheinander aus: Zuerst werden alle (n) für eine bestimmte Abstandsklasse ermittelten Werte E_i bzw., im Falle von Mittelwertbildung (siehe letztes Kapitel), $\bar{E}_j$ nach der Größe sortiert so dass gilt:

$$E_{i-1} < E_i < E_{i+1}$$

(bei Mittelwertbildung) $\quad \bar{E}_{j-1} < \bar{E}_j < \bar{E}_{j+1}$

Es wird nun angenommen, dass elf aufeinanderfolgende Werte E_i bzw. $\bar{E}_j$ mit der Eigenschaft

$$E_{i-1} < E_i < E_{i+1}$$

(bei Mittelwertbildung) $\quad \bar{E}_{j-1} < \bar{E}_j < \bar{E}_{j+1}$

auf einer Gaußkurve mit dem Mittelwert

$$< \overline{E_i} >= \frac{1}{11} \sum_{a=-5}^{5} E_{i+a} \tag{3.1}$$

(bei Mittelwertbildung) $\quad < \overline{E_j} >= \frac{1}{11} \sum_{a=-5}^{5} \overline{E}_{j+a}\, g_{j+a}$ (3.2)

und der Varianz

$$\sigma_i^2 = \frac{1}{10} \sum_{a=-5}^{5} (E_{i+a} - < \overline{E_i} >)^2 + \sigma_0^2 \tag{3.3}$$

(bei Mittelwertbildung) $\quad \sigma_j^2 = \frac{1}{10} \sum_{a=-5}^{5} (\overline{E}_{j+a} - < \overline{E_j} >) g_{j+a})^2 + \sigma_0^2$ (3.4)

(g_{j+a} ist hierbei die Anzahl der in einem Mittelwert $\bar{E}_{j+a}$ zusammengefassten Werte)

normalverteilt sind. Jetzt kann die Wahrscheinlichkeitsdichteverteilung durch eine Summe normierter Gaußfunktionen angenähert werden durch:

$$p(\chi \mid WP_\lambda) = \frac{1-p_0}{J} \sum_{i=1}^{J} \frac{1}{\sigma_i \sqrt{2\pi}} e\left(-\frac{1}{2} \left(\frac{\chi - < \overline{E}_i >}{\sigma_i} \right)^2 \right) + p_0 \qquad (3.5)$$

$$p(\chi \mid WP_\lambda) = \frac{1-p_0}{J} \sum_{j=1}^{J} \frac{1}{\sigma_j \sqrt{2\pi}} e\left(-\frac{1}{2} \left(\frac{\chi - < \overline{E}_j >}{\sigma_j} \right)^2 \right) g_j + p_0 \qquad (3.6)$$

(bei Mittelwertbildung)

$p(\chi|WP_\lambda)$ ist hierbei die Wahrscheinlichkeitsdichte für einen bestimmten Abstand oder Signalvolumen χ, die ein gegebenes Wasserstoffatompaar WP_λ (Zuordnungsmöglichkeit für eine bestimmtes NOESY-Signal) besitzt. Die Werte χ durchlaufen bei der Erzeugung von Abstandswahrscheinlichkeitsdichteverteilungen einen Abstandsbereich von 1,7 Å –100Å mit jeweils einer Schrittweite von 0,05 Å. Bei Volumenwahrscheinlichkeitsdichteverteilungen wird χ für einen Wertebereich von 0 $Å^3$ –0,042$Å^3$ mit einer Schrittweite von 0,0000042 $Å^3$ berechnet. Dies entspricht einen Abstandsbereich ausgehend vom Unendlichen bis 1,7 Å. J ist die Anzahl der für ein bestimmtes Wasserstoffatompaar WP_λ gefundenen Atomabstände bzw. Signalvolumina. Bei Zusammenfassung von Werten entspricht J der Anzahl der resultierenden Mittelwerte. Zur Vermeidung von lokalen Extremwerten aufgrund von verschwindend kleinen Varianzen wird ein zusätzlicher Wert σ_0^2 zur Varianz addiert, was für eine von Null verschiedene Mindestbreite der entsprechenden Gaußkurve sorgt. Als Wert für σ_0 wurde die jeweilige Datenauflösung (Schrittweite der χ Werte) der Wahrscheinlichkeitsdichteverteilung eingesetzt. Der Grund für die Addition dieses zusätzlichen Wertes liegt darin, dass die Werte der Varianzen in diesen Zusammenhang extrem klein bzw. Null werden können. Dies kann beispielsweise dann geschehen, wenn mehrere hintereinander liegende Werte von Messdaten sich betragsmäßig sehr wenig voneinander unterscheiden oder, aufgrund von rechnerinterner Rundungsverfahren, sogar identisch werden. Bei Volumenwahrscheinlichkeitsdichteverteilungen beträgt σ_0 =0,0000042 bzw. bei Abstandswahrscheinlichkeitsdichteverteilungen ist σ_0 =0,05, was den jeweiligen Datenauflösungen der Verteilungen entspricht. Die für die Wahrscheinlichkeitsdichteverteilungen jeweils berechneten Wahrscheinlichkeitsdichten p sind mit einer Genauigkeit von sechs Stellen hinter dem Komma abgespeichert. Um einen

Nulleintrag im Falle von $p < 10^{-6}$ zu vermeiden, wird zu jeder berechneten Wahrscheinlichkeitsdichte p eine Grundwahrscheinlichkeitsdichte $p_0 = 10^{-6}$ hinzuaddiert.

3.5.2.4 Abspeicherung der Verteilungskurven

Die berechneten Werte einer bestimmten Wahrscheinlichkeitsdichteverteilung werden jeweils in einer eigenen Datei sequentiell abgespeichert. Am Anfang der Datei sind noch zusätzliche Informationen, wie die Anzahl und der Mittelwert aller innerhalb des Strukturdatensatzes gefundenen Abstände sowie jeweils der größte und der kleinste ermittelte Abstand vermerkt. Der Name einer solchen Datei entspricht jeweils der Abstandklasse für welche die betreffende Wahrscheinlichkeitsdichteverteilung berechnet wurden (z.B. Ala_HA-Arg_HB i+3 *(aus Datenbank 3),* [Ala, Arg, i+3] *(aus Datenbank 1*), [HA-HB i+3] *(aus Datenbank 2)).* Dies vereinfacht den programmtechnischen Zugriff auf die jeweils benötigten Dateien während der Ausführung des Programms *KNOWNOE* erheblich.

```
DISTANCES_FOUND [N]->978
DISTANCE_MAX [A]   ->10.592770
DISTANCE_MIN [A]   ->1.976063
DISTANCE_MID [A]   ->5.534746
35499.195775
43909.151317
25129.022276
9000.776088
3913.877026
2902.537240
2304.486882
```

Abbildung 3.13: Ausschnitt vom Dateiinhalt einer Wahrscheinlichkeitsdichteverteilung

Bei der vorhergehenden Programmversion kam ein programmtechnisch recht umständliches Verfahren für Ermittelung der jeweils benötigten Datei bzw. deren Dateinamen zur Anwendung. Dies führte zur starken Aufblähung des Programmcodes. Durch die hier angewandte Weise der Dateinamensgebung konnte der Programmcode um etwa 70 % reduziert werden. Abbildung 3.13 zeigt einen Ausschnitt aus einer Datei für eine Wahrscheinlichkeitsdichteverteilung. Jeder Wert (Wahrscheinlichkeitsdichte) einer Verteilung ist mit einer Genauigkeit von maximal 6 Stellen hinter dem Komma abgespeichert.

Die Werte für Volumenwahrscheinlichkeitsdichteverteilungen sind jeweils für den Bereich von 0 bis 0,042 $Å^3$ (entspricht: ∞ Å - 1,7 Å) mit einer Schrittweite von je 4.2 x 10^{-6} $Å^3$ abgespeichert. Die y-Werte der Abstandswahrscheinlichkeitsdichteverteilungen sind jeweils für einen Abstandsbereich von 1,7 –100 Å mit einer Schrittweite von je 0,05 Å angegeben. Die erzeugten Wahrscheinlichkeitsdichteverteilungen wurden in Abhängigkeit ihrer entsprechenden Abstandsklasse auf jeweils unterschiedliche Ordner aufgeteilt. Dabei wurden jeweils gesonderte Ordner (i+0, i+1, i+2, i+3, i+4, i+5, i+6, i+7, i+8, i>8) für Verteilungen von Abstandklassen mit jeweils gleichen relativen Sequenzabständen ΔS angelegt. Bei der *Datenbank 3* enthalten diese Ordner noch einmal 20 Unterordner, wobei jeder den Namen einer der 20 natürlich vorkommenden Aminosäuren im Dreibuchstabencode trägt. In jedem dieser Ordner befinden sich jeweils diejenigen Verteilungen, deren Namen mit dem entsprechenden Aminosäurenamen beginnt. In Abbildung 3.14 ist die Ordnerstruktur der *Datenbank 3* (s. Kap. 4.1.1.2) grafisch dargestellt.

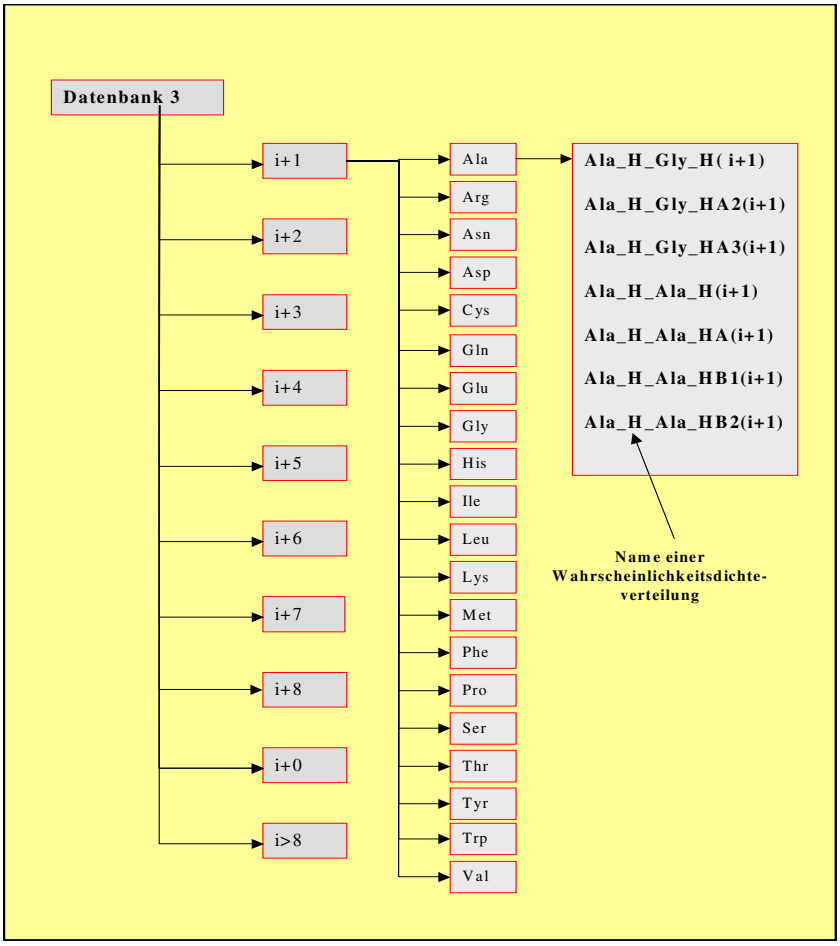


Abbildung 3.14: Aufbau der Datenbank 3. In der Grafik ist die Ordnerstruktur der *Datenbank 3* dargestellt .Für den Ordner i+1 ist die weitergehende Ordner- bzw. Dateistruktur im Detail aufgeführt. Diese ist für die restlichen Ordner analog.

3.5.2.5 Reduzierung des Speicherbedarfes der erweiterten Datenbanken

Die im Rahmen dieser Arbeit erzeugten Datenbanken aus Wahrscheinlichkeitsdichteverteilungen besitzen mit einem Speicherbedarf von mehr als 21 GB auch für den heutigen Stand der Technik eine recht unhandliche Größe. Somit dürfte die Bereitschaft eines Benutzers mehrere Gigabyte Speicherplatz für eine Softwareanwendung auf seinen PC frei zu stellen sehr gering sein. Deshalb war es nötig ein Verfahren zu finden die vorhandene Datenmenge zu reduzieren. Das benötigte Verfahren musste folgende Eigenschaften aufweisen:

1. Starke Reduktion des benötigten Speicherbedarfs der neuen Datenbanken ohne Verlust wesentlicher Informationen.
2. Das Verfahren muss große Datenmengen in kurzer Zeit automatisch umwandeln können.
3. Der ursprüngliche Verlauf einer bestimmten Wahrscheinlichkeitsdichteverteilung sollte bei Bedarf aus den reduzierten Daten möglichst schnell wiederhergestellt werden können.

Es wurden zunächst drei mögliche Verfahren in Betracht gezogen:

1. Anwendung eines der gängigen Datenkomprimierungsprogramme wie z.B. *WinZip [101]*.
2. Ersetzen der Wahrscheinlichkeitsdichteverteilungen durch Polynome n-ten Grades.
3. Anwendung der kubischen Splineinterpolation

Bei der ersten und einfachsten Möglichkeit lassen sich die Daten um mehr als 90% reduzieren (von etwa 100kB auf 8kB pro Verteilung). Der Nachteil dieses Verfahren besteht darin, dass bei Bedarf einer bestimmten Verteilung diese zuerst dekomprimiert und nach der Benutzung wieder komprimiert werden muss. Dies würde zu erheblichen Zeitverzögerungen während der Anwendung (*KNOWNOE*) führen. Außerdem wäre das Programm *KNOWNOE* und somit auch das Softwarepaket *AUREMOL* abhängig vom Vorhandensein externer Programme. Deshalb schied diese Methode aus. Die Lösung bzw. Aufstellung der Gleichungssysteme zur Berechnung der benötigten Koeffizienten der jeweils gesuchten Polynome ist aufgrund der oft großen Anzahl von Minima bzw. Maxima der Verteilungskurven praktisch meist nicht durchführbar. Erschwerend kommt hinzu, dass der Ordnungsgrad der gesuchten Funktionen nicht bekannt ist. Die Anwendung der kubischen Splineinterpolation hingegen schien für die

vorliegende Aufgabenstellung hinsichtlich der Durchführbarkeit am geeignetsten. Mit diesen Verfahren wurde der Speicherbedarf der *Datenbank 3* von etwa 7 GB auf etwa 200 MB, ohne nennenswerte Veränderung bezüglich der Zuordnungsqualität, reduziert.

3.5.2.5.1 Anwendung kubischer Interpolationssplines

Die Grundidee des hier angewandten Verfahren zur Datenreduktion liegt darin, dass man die ursprünglichen Dateien aus Wahrscheinlichkeitsdichteverteilungen durch Dateien (*Knotenpunktdateien*) ersetzt, in denen sich jeweils nur eine begrenzte Anzahl ausgesuchter Datenpunkte (*Knotenpunkte*) der ursprünglichen Verteilungskurven befinden (Abb. 3.15). Wird nun eine bestimmte Verteilung z.B. während der Anwendung benötigt, kann man die ursprüngliche Kurve durch Verbinden der vorhandenen *Knotenpunkte* X_i mit Polynomen 3.Grades *(kubische Interpolationssplines* [44]) rekonstruieren.

$$S(x) = S_i(x) = a_i + b_i(x - X_i) + c_i(x - X_i)^2 + d_i(x - X_i)^3 \qquad (3.7)$$
$$mit:$$
$$x \in [X_i, X_{i+1}]$$
$$(i = 1,2,...,N-1)$$

Die Verbindungskurven zwischen zwei Knotenpunkten X_i und X_{i+1} haben dabei in etwa den Verlauf eines dünnen biegsamen Lineals (englisch *spline),* welches man durch zwei fixierte Punkte legt. Die abschnittsweise Annäherung einer Kurve durch *kubische Interpolationssplines* hat den zusätzlichen Vorteil, dass sich mögliche Oszillationen weitgehend vermeiden lassen. Für die Extraktion geeigneter Knotenpunkte aus einem bestimmten Funktionsgraphen wurde das Programm *Reduce_Prob_Tabs.c* geschrieben, auf welches im folgenden Kapitel näher eingegangen wird.

0	***0.000000***
6	***0.017717***
12	***0.840263***
19	***0.171622***
26	***0.060427***
33	***0.159531***
40	***1.580866***

Abbildung 3.15: Ausschnitt aus einer Knotenpunktdatei. Zu sehen ist eine Liste ausgesuchter Datenpunkte (x/y Werte) einer Volumenwahrscheinlichkeitsdichtverteilung [HA/HN(i+1)]. Der x -Wert (linke Spalte) liegt, aus Gründen der Speicherplatzersparnis, als ganze Zahl vor. Um den ursprünglichen x-Wert zu erhalten, muss dieser Wert mit 4.2×10^{-5} multipliziert werden.

3.5.2.5.2 Automatische Bestimmung geeigneter Knotenpunkte

Zur automatischen Bestimmung geeigneter *Knotenpunkte* aus einem gegebenen Funktionsgraphen wurde das Programm *Reduce_Prob_Tabs.c* geschrieben. Geeignete *Knotenpunkte* sind insbesondere Datenpunkte, welche für das Profil eines bestimmten Funktionsgraphen prägend sind. Das Programm *Reduce_Prob_Tabs.c* ist dazu ausgelegt eine große Anzahl von Wahrscheinlichkeitsdichteverteilungen bzw. Funktionsgraphen automatisch hintereinander zu bearbeiten. Dem Programm muss man hierfür nur eine Dateinamensliste der zu bearbeitenden Kurven übergeben. Diese werden dann automatisch hintereinander eingelesen und bearbeitet. Das Programm *Reduce_Prob_Tabs.c* ist für die Bearbeitung von Funktionsgraphen mit n Datenpunkten $DP_i = \{ x_i, y_i \}$ $(i=1,2..n)$ folgender Eigenschaften ausgelegt:

1. $x_{i-1} < x_i < x_{i+1}$
2. $x_i - x_{i-1} = x_{i+1} - x_i$
3. *alle* $x_i \geq 0$
4. *alle* $y_i > 0$

Die dem Programm übergebenen Datenpunkte sind in diesem Falle die Kurvenpunkte, der im Rahmen der Arbeit generierten Wahrscheinlichkeitsdichteverteilungen. Die Zusammenstellung der benötigten Knotenpunkte aus einer gegebenen Kurve erfolgte jeweils in einem iterativen Prozess (Abb. 3.18). Das Programm führt hierbei folgende Arbeitschritte aus:

1. *Einlesen der Datenpunkte eines in der übergebenen Liste stehenden Funktionsgraphen (= Wahrscheinlichkeitsdichteverteilung).*

2. *Ermittlung wichtiger Knotenpunkte.* Zunächst wird eine Kurvenanalyse durchgeführt. Dabei werden folgende Datenpunkte (Knotenpunkte) des zu bearbeitenden Funktionsgraphen ermittelt:
   *Erster und letzter Datenpunkt der Kurve
   *Maxima
   *Minima
   *Wendepunkte
   *Scheitelpunkt

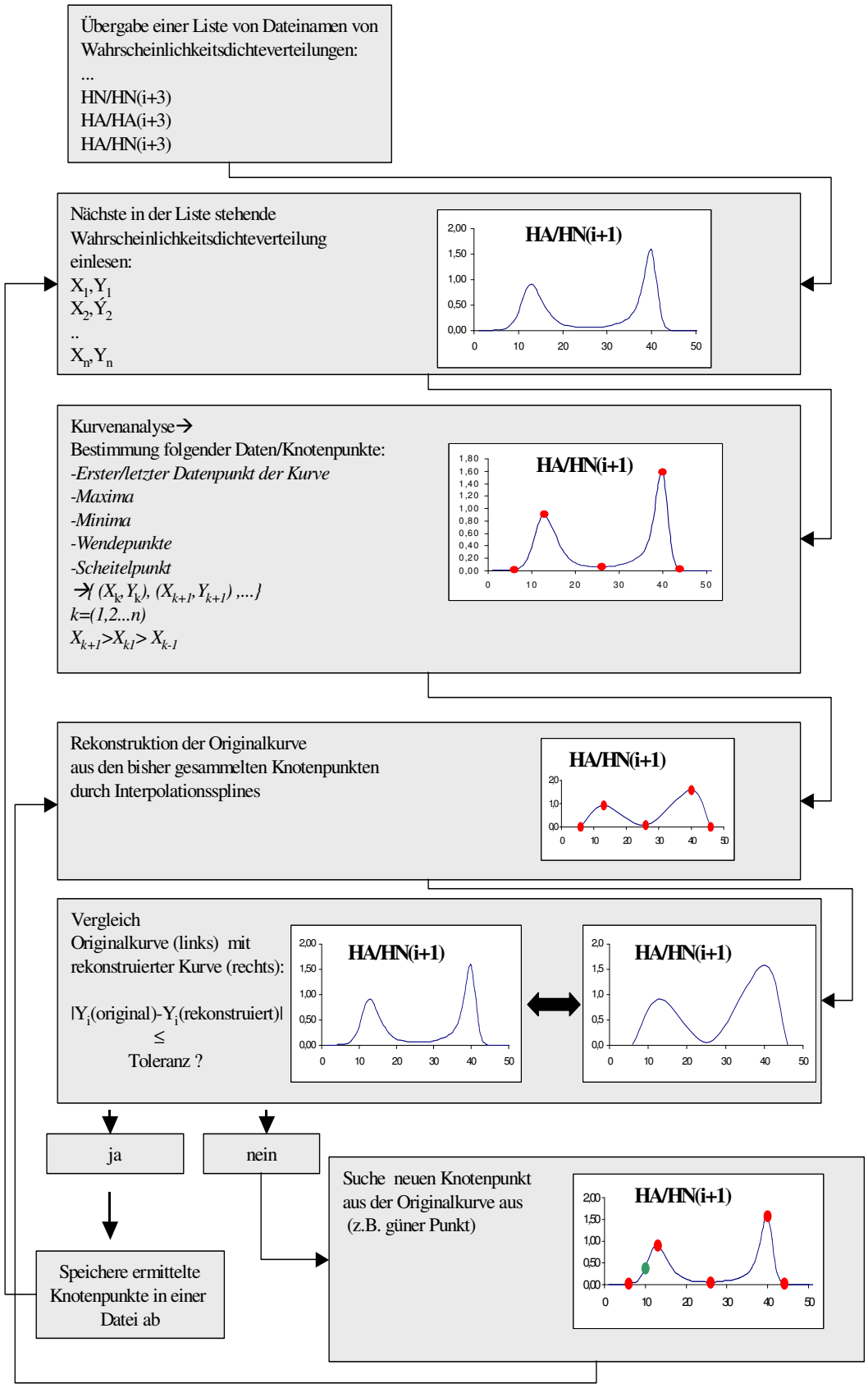


Abbildung 3.16: Programmablauf vom Programm *Reduce_Prob_Tabs..c*

In Abbildung 3.17 sind die angewandten Kriterien zur Identifikation der genannten Punkte aufgeführt.

3. *Rekonstruktion der Originalkurve.* Aus dem im 2. Schritt ermittelten Knotenpunkten KP_i (X_i, Y_i) wird zunächst versucht, die Originalkurve durch Verbinden mit *kubischen Interpolationsplines* zu rekonstruieren. Für die Berechung der *Interpolationssplines Si(x)* wurden die optimierten Programmfunktionen *splint* [46] bzw. *spline* [46] in das Programm integriert.Zuerst wird Funktion *spline* ausgeführt, die an jedem ermitteltem *Knotenpunkt* $KP(X_i, Y_i)$ die 2. Ableitung $S''(X_i)$ $(i=1,2,..N)$ ermittelt und in einer Liste (*Array*) vermerkt.

Art des ausgezeichneten Punktes	**Identifikationskriterien**
Maxima	$y_{i-1} < y_i > y_{i+1}$
Minima	$y_{i-1} > y_i < y_{i+1}$
Wendepunkte **Kurve aufsteigend:** **$f'(x_i)>0$**	$y_{i-2} < y_{i-1} < y_i < y_{i+1} < y_{i+2}$ $\lvert(y_{i-2} - y_{i-1})\rvert > \lvert(y_{i-1} - y_i)\rvert$ $\lvert(y_i - y_{i+1})\rvert < \lvert(y_{i+1} - y_{i+2})\rvert$
Wendepunkte **Kurve absteigend:** **$f'(x_i)<0$**	$y_{i-2} > y_{i-1} > y_i > y_{i+1} > y_{i+2}$ $\lvert(y_{i-2} - y_{i-1})\rvert > \lvert(y_{i-1} - y_i)\rvert$ $\lvert(y_i - y_{i+1})\rvert < \lvert(y_{i+1} - y_{i+2})\rvert$
Scheitelpunkt **Kurve absteigend:** **$f'(x_i)<0$)**	$(y_{i-1} - y_i) / (y_i - y_{i+1}) = Maximal$

Abbildung 3.17. Identifikationskriterien für ausgezeichnete Punkte einer Kurve. In der Tabelle sind die Kriterien aufgeführt, die ein Datenpunkt auf einer gegebenen Kurve zugleich erfüllen muss, um vom Programm *Reduce_Prob_Tabs.c* als einer der hier aufgeführten ausgezeichneten Punkte identifiziert zu werden.

Die ermittelten Datenpunkte werden nun in einer Liste für die gesammelten Knotenpunkte $KP_i(X_i,Y_i)$ $(i=1,2,..N)$ vermerkt. Für die Knotenpunkte innerhalb der Liste gilt:

$$X_{i-1}<X_i<X_{i+1}$$

Der Funktion müssen folgende Eingangsinformationen übergeben werden:

1. Liste aller x-Werte X_i der ermittelten Knotenpunkte mit: $X_{i-1}<X_i<X_{i+1}$.
2. Liste aller y-Werte Y_i der ermittelten Knotenpunkte.
3. Anzahl N aller ermittelten Knotenpunkte.
4. Betrag der ersten Ableitung am ersten und letzten Knotenpunkt.

Es wird davon ausgegangen, dass die Werte der Kurven außerhalb des angegeben Wertebereiches keinen wesentlichen Änderungen mehr unterliegen bzw. gegen Null gehen. Deshalb wurde die erste Ableitung des ersten bzw. letzten Knotenpunktes (= erster bzw. letzter Datenpunkt der Kurve) jeweils auf den Wert Null gesetzt. Nachdem die zweiten Abeitungen für die vorhandenen Knotenpunkte berechnet wurden, erfolgt die eigentliche Rekonstruktion der Originalkurve mit der Funktion *splint.* Die Funktion berechnet für einen gegebenen x-Wert den zugehörigen y-Wert aus. Zuerst muss die Funktion den entsprechenden *Interpolationspline bzw.* seine Koeffizienten ermitteln. Diese sind abhängig von den beiden Knotenpunkten zwischen denen der in Frage stehende x-Wert jeweils liegt. Der x-Wert muss dabei innerhalb des Bereiches zwischen den ersten bzw. den letzten vorhandenen Knotenpunkt liegen ($X_1 \le x \le X_N$). Folgende Informationen müssen der Funktion *splint* vor ihrer Ausführung übergeben werden:

1. Liste aller x-Werte X_i der vorhandenen Knotenpunkte mit $X_{i-1}<X_i<X_{i+1}$.
2. Liste aller y-Werte Y_i der vorhandenen Knotenpunkte.
3. Liste der mit der Funktion *spline* ermittelten 2. Ableitungen an den Knotenpunkten.
4. Anzahl N aller vorhandenen Knotenpunkte.
5. x-Wert für den der y-Wert berechnet werden soll.

Um die rekonstruierte Kurve mit der Originalkurve später vergleichen zu können, wird diese mit der Datenauflösung der Originalkurve erstellt.

4. *Vergleich der rekonstruierten Kurve mit der Originalkurve.* Hier wird die intern rekonstruierte Kurve mit der Originalkurve Punkt für Punkt verglichen. Dabei darf der Absolutbetrag der Differenz der entsprechenden y-Werte einen bestimmten definierten Toleranzwert nicht überschreiten. Es gilt:

$|y_i \text{ (original)} - y_i \text{ (rekonstruiert)}| \leq \text{Toleranz} \quad \text{mit } i=\{1,2,..n\}$

Der Toleranzwert wurde hier auf 1/10 000 des Maximalwertes der Originalkurve gesetzt.
Die Wahl dieses Qualitätskriteriums bewirkt, dass die rekonstruierte Kurve in Bereichen mit relativen großen bzw. statistisch signifikanten Werten eine relativ hohe Übereinstimmung mit der Originalkurve haben muss. In Bereichen mit relativ zum Maximum kleinen Werten hingegen, lässt dieser Toleranzwert größere Unterschiede zu. Dies hat den Vorteil, dass in diesen Bereichen weniger Knotenpunkte ausgewählt werden müssen. Es wird davon ausgegangen, dass Werte, die wesentlich kleiner als der Maximalwert sind, statistisch nicht oder kaum relevant sind. Falls ein Datenpunkt der rekonstruierten Kurve die erforderliche Genauigkeit nicht erreicht, wird Arbeitschritt 5. ausgeführt. Ansonsten wird zu Schritt 7. übergegangen.

5. Hinzufügen zusätzlicher Datenpunkte. Als neuer Knotenpunkt KP_{neu} wird dabei derjenige Datenpunkt der Originalkurve in die Liste der bereits gesammelten Knotenpunkte hinzugefügt, welcher genau in der Mitte zwischen den Knotenpunkten X_i und X_{i+1} liegt in denen die rekonstruierte Kurve nicht der geforderten Genauigkeit entspricht:

$KP_{neu} = (X_{neu}, Y_{neu})$

mit

$X_{neu} = X_i + (X_{i+1} - X_i) / 2$

$Y_{neu} = y\text{-Wert bei } X_{neu} \text{ der Originalkurve}$

6. *Gehe zurück zu Arbeitschritt 3*

7. *Abspeicherung der gesammelten Knotenpunkte*

Die gesammelten Knotenpunkte für einen bestimmten Funktionsgraphen werden jeweils in einer separaten Datei abgespeichert. Um weiter Speicherplatz zu sparen, wurden anstatt der x-Werte der Knotenpunkte der Quotient aus x-Wert und Datenauflösung der Originalkurve

abgespeichert. Dieser Quotient ergibt dabei immer eine ganze Zahl (*Integerzahl*), welche weniger Speicherplatz im Vergleich eines Bruches (*Flieskommazahl*) beansprucht.

8. *Gehe zurück zu Punkt 1(=nächste Datei einlesen)*

Abbildung 3.18 zeigt anhand einer Abstandswahrscheinlichkeitsdichtverteilung für die Abstandsklasse [HN, HA (i+1)], wie mit steigender Anzahl eingesetzter Knotenpunkte die Qualität bzw. Ähnlichkeit der rekonstruierten Kurve mit der Originalkurve zunimmt. Das Hauptproblem der Anwendung von *Interpolationssplines* ist das Auftreten von Oszillationen, was zu unlässigen negativen Werten bei der Kurvenekonstruktion führen kann (s. Abbildung 4.25). Oszillationen treten vor allem bei zu großen Abständen zwischen zwei *Knotenpunkten* auf. Da sich negative Werte auch bei noch so kleinen eingestellten Toleranzwerten oder einer entsprechend hohen Knotenpunktdichte nicht ausschließen lassen, wurden diese durch die Grundwahrscheinlichkeitsdichte $p_0 = 10^{-6}$ (s. Kap. 3.5.2.3) ersetzt.

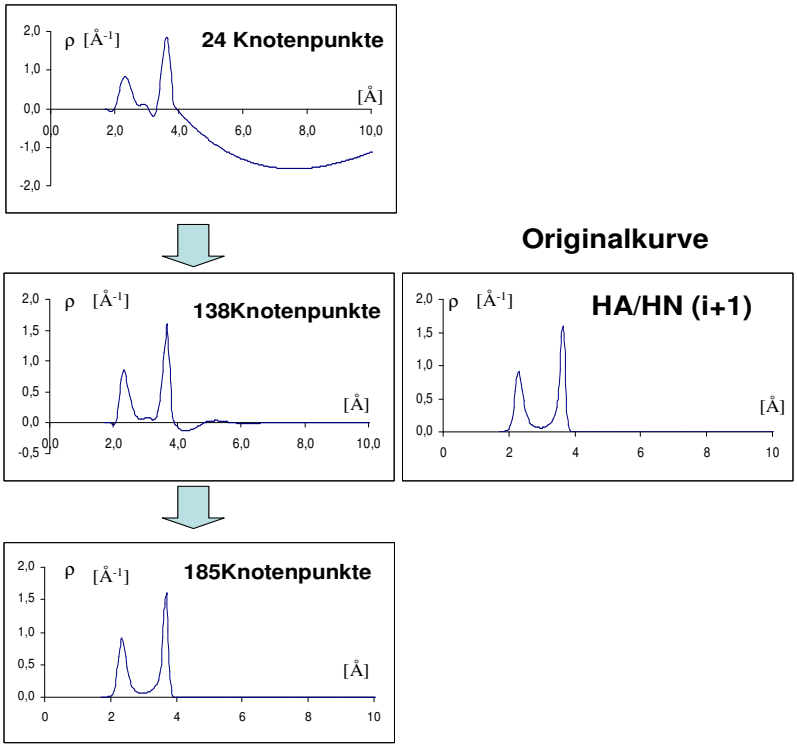


Abbildung 3.18 Rekonstruktion einer Wahrscheinlichkeitsdichteverteilung [HA,HN (i+1)] (rechts) mit Hilfe der *kubischen Splineinterpolation.* Die Grafiken (links) zeigen drei rekonstruierter Kurven auf der Basis von jeweils unterschiedlicher Anzahl benutzter *Knotenpunkte.*

3.6 Testreihen mit dem Programm *KNOWNOE*

3.6.1 Prinzipielle Vorgehensweise

Bei den hier durchgeführten Testreihen wurden jeweils unter Benutzung unterschiedlicher Gruppen von Wahrscheinlichkeitsdichteverteilungen (frühere Datenbank und neue Datenbanken 1-3) und verschiedenen Versuchbedingungen Zuordnungen der simulierten 2D-NOESY-NMR-Spektren der Proteine HPr und *Tm*CSP mit dem Programm *KNOWNOE* erstellt. Die Testreihen wurden unter verschiedenen eingestellten Suchradien durchgeführt, da sich die Eigenschaften der Zuordnungsmöglichkeiten für die NOESY-Signale im Abhängigkeit vom Suchradius entsprechend stark ändern (s. Kap. 4.2.1.2). Vor jedem Programmablauf wurden alle bereits vorhandenen Zuordnungen mithilfe einer bereits vorhandenen Funktion im Programm *AUREMOL* aus der *Masterliste* entfernt. Die mit dem Programm *KNOWNOE* erstellten Zuordnungen wurden dann im Anschluss mit Hilfe des Programms *Analyse_KNOWNOE_Result.c* (s. Kap. 3.6.3) automatisch ausgewertet. Die Analyse beruht hierbei auf einen Vergleich der erstellten Signalzuordnungen mit den Signalzuordnungen innerhalb der ursprünglichen rückgerechneten NMR-Spektren.

3.6.2 Allgemeine Versuchbedingungen

Im folgendem werden die während der Durchführung der Versuchsreihen konstanten Bedingungen im Programm *KNOWNOE* aufgeführt. Diese währen:

1. *Die vom Benutzer definierte Wahrscheinlichkeitsgrenze.* Sie wurde meist bei fast allen Versuchen auf p=0,98 gesetzt (Ausnahme: s. Kap. 4.2.2.8).

2. *Die Toleranz der chemischen Verschiebung.* Mit der Wahl dieses Wertes sollte erreicht werden, dass man auch bei jeweils verschiedenen eingestellten Suchradien eine genügend große Anzahl von zwei- und dreideutig zugeordneten NOESY-Signalen erhält um eine brauchbare statistische Auswertung durchführen zu können. Als Orientierung dienten hierbei die Ergebnisse aus Kapitel 4.2.1.1. Die eingestellte Toleranz der chemischen Verschiebung betrug bei den Testreihen immer 0,015 ppm.

3. *Der Kalibrierungsfaktor.* Für die Berechnung der Kalibrierungsfaktoren erschien es sinnvoll ein mittelstarkes NOESY-Signal, dass einem Abstand von etwa 3 Å entspricht, zu wählen [13]. Für das 2D-NOESY-NMR-Spektrum des Protein *Tm*CSP wurde hierbei das NOESY-Signal zwischen dem *HA*-Atom der Aminosäure Glycin an der Sequenzposition 15 und dem *HN*-Atom der Aminosäure Threonin an der Sequenzposition 18 benutzt (Abstand=3.105Å). Für das 2D-NOESY-NMR-Spektrum des Proteins HPr wurde das intraresiduale NOESY-Signal (Abstand=3.012Å) zwischen dem *HA2*-Atom und dem *HN*-Atom der Aminosäure Glycin an der Sequenzposition 13 verwendet. Im Zuge der Weiterentwicklung von *AUREMOL* wird der Kalibrierungsfaktor allerdings seit neueren automatisch über das Programm *RELAX* bestimmt.

4. *Reduktion der Datenauflösung bei der Zuordnung dreideutiger NOESY-Signale.* Bei der Berechnung der wahrscheinlichsten Zuordnung von dreideutigen NOESY-Signalen wurde die Datenauflösung der jeweils benötigten Wahrscheinlichkeitsdichteverteilungen vor der Benutzung um den Faktor 10 reduziert. Hierbei wurde der Mittelwert aus jeweils 10 hintereinanderliegenden Datenpunkten gebildet und aus den resultierenden Werten intern eine temporäre Verteilung erzeugt. Es hat sich nämlich gezeigt, dass bei einer größeren Datenauflösung die Rechenzeit in unakzeptabler Weise ansteigen (mehrere Stunden) würde. Der Grund dafür liegt darin, dass die Ausdrücke der Integrale zur Berechnung der benötigten Wahrscheinlichkeiten für dreideutige Zuordnungen komplexer bzw. im höheren Grade verschachtelt sind.

3.6.3 Analyse automatisch zugeordneter NOESY - Signale

Wie bereits erwähnt, wurde das Programm *Analyse_KNOWNOE_Result.c* zur Analyse der mit dem Programm *KNOWNOE* erstellten NOESY-Signalzuordnungen benutzt. Es ist etwa 600 Zeilen lang. Dem Programm müssen folgende Dateien übergeben werden:

1. Die vollständig zugeordnete *Masterliste* des jeweils simulierten 2D-NOESY-NMR-Spektrums.

2. Die vom Programm *KNOWNOE* zugeordnete *Masterliste* des simulierten 2D-NOESY-NMR-Spektrums.

3. Die vom Programm *KNOWNOE* erstellte Ausgabedatei *output.txt* (s. Kap. 2.5). Sie enthält noch zusätzliche wichtige Informationen wie z.B. die Anzahl der sich ergebenen Zuordnungsmöglichkeiten für ein bestimmtes NOESY-Signal.

Das Programm *Analyse_KNOWNOE_Result.c* vergleicht die Zuordnung jedes Signals aus dem rückgerechneten Spektrums mit derjenigen Zuordnung, die das Programm *KNOWNOE* für das Signal erstellt hat. Sind beide Zuordnungen eines bestimmten Signals identisch, wird die vom Programm *KNOWNOE* erstellte Zuordnung als richtig zugeordnet deklariert. Bei Nichtübereinstimmung wird die Zuordnung entsprechend als falsch bewertet. Vor der Ausführung des Programms *KNOWNOE* müssen alle bereits vorhandenen Zuordnungen des simulierten NMR-Spektrums (Kopie vom Originalspektrum) entfernt werden. Ein Großteil der erhaltenen Ergebnisse beruht auf den von diesem Programm erstellten Analysen. Das Programm kann separat für alle oder verschiedene ausgewählte Gruppen von NOESY-Signalen Analysen bezüglich ihrer Häufigkeit, Zuordnungsanzahl und Richtigkeit der erstellten Zuordnungen erstellen.

4.Ergebnisse

4.1 Aufbau umfangreicher Datenbanken aus Wahrscheinlichkeitsdichteverteilungen

In diesem Kapitel soll der Inhalt der in dieser Arbeit erzeugten Datenbanken aus Abstand- und Volumenwahrscheinlichkeitsdichteverteilungen vorgestellt werden.
Wie bereits in Kapitel 2.5.3 beschrieben, bilden bei der Berechnung der wahrscheinlichsten Zuordnungsmöglichkeit mehrdeutiger NOESY-Signale Volumenwahrscheinlichkeitsdichte-verteilungen die wesentliche Informationsgrundlage. Für ein optimales Ergebnis ist es wichtig, dass die benutzte Datenbank für alle möglichen Zuordnungen bzw. Atompaare eine entsprechend repräsentative Volumenwahrscheinlichkeitsverteilung enthält. Ziel vom Algorithmus vom Programm *KNOWNOE* ist es, möglichst vielen zwei- und dreideutigen NOESY-Signalen dasjenige Atompaar mit jeweils einer hoher Wahrscheinlichkeit zuzuweisen, welches mindestens 90% des Signalvolumens erklärt.
Unter der Benutzung des bereits vorhandenen Satzes von Wahrscheinlichkeitsverteilungen ist sowohl die erreichte Anzahl von Zuordnungen sowie deren Sicherheit recht unbefriedigend gewesen. Kernziel war es, durch Einsatz eines qualitativ hochwertigeren Datensatzes aus Wahrscheinlichkeitsverteilungen die Zuordnungsqualität mehrdeutiger NOESY-Signale entsprechend stark zu verbessern. Dies sollte durch eine höhere Anzahl von eingesetzten und zugleich nichtredundanten Proteinstrukturen als Datenbasis zur Erzeugung von Verteilungen, durch eine erhöhte Anzahl zu Verfügung stehender Verteilungen und einem akkurateren mathematischen Verfahren (Summierung über Gaußkurven) zur Berechnung einer Verteilungskurve erreicht werden. Darüber hinaus stellen die hier erzeugten Datenbanken eine wertvolle Quelle struktureller Information dar, welche für die Optimierung vieler anderer für die Strukturbestimmung benötigter Arbeitschritte eingesetzt werden kann (s. Kap. 6.2)

4.1.1 Eigenschaften der erweiterten Datenbank

Hier wird auf die Eigenschaften der neuen Wahrscheinlichkeitsdichteverteilungen näher eingegangen.

4.1.1.1 Unterschiede zur früheren Datenbank

Die im Rahmen der Arbeit erstellten Datenbanken aus Wahrscheinlichkeitsdichte-verteilungen unterscheiden sich von den früheren Verteilungen [53] in folgenden wesentlichen Punkten:

1. Anzahl der Proteinstrukturen als Datenbasis.

2. Eigenschaften der benutzten Proteinstrukturen.

3. Angewandtes mathematische Verfahren zur Berechnung einer Verteilungskurve.

4. Art der Verteilungen.

5. Datenauflösung einer Verteilung.

6. Bildung von Abstandsklassen.

Durch den Einsatz einer größeren Anzahl von Proteinstrukturen (1107 anstatt 326) als Datenbasis, steht eine entsprechend größere Anzahl von Atomabständen für die Erzeugung einer Wahrscheinlichkeitsdichteverteilung zur Verfügung. Dies soll zu einer verbesserten statistische Aussagekraft einer Verteilung führen (1). Damit die neuen Verteilungen möglichst repräsentativ im Bezug auf unterschiedliche Proteine sind, beträgt die paarweise Sequenzidentität zwischen den Strukturen innerhalb des gegebenen Proteinstrukturdatensatzes untereinander weniger als 25%. Dies soll dazu führen, dass die hier angewandte Zuordnungsmethode für verschiedene zu untersuchende Proteine gleich gut funktioniert (2). Die Berechnung von Volumenwahrscheinlichkeitsdichteverteilungen aus Atomabständen wurde mittels Summation über Gaußkurven [100] durchgeführt (s. Kap. 3.5.2.3). Hierbei wurde zuvor jeder für eine bestimmte Verteilung ermittelte Abstand r[Å] mit dem Ausdruck $V=(1/r^6)*$ Å^9 in ein entsprechendes Signalvolumen umgewandelt. Das Verfahren hat den Vorteil, dass sogenannte „*statistische Ausreißer*“ innerhalb einer gegebenen Wertemenge weitgehend weggemittelt werden und somit den Kurvenverlauf in nur geringen Maß verzerren können. Zusätzlich ist das Verfahren in der Lage statistisch schlecht definierte Wertebereiche, aufgrund mangelnder oder fehlender Werte, partiell zu kompensieren (3). Die neuen

Verteilungen wurden hier als Wahrscheinlichkeitsdichteverteilungen, anstatt wie vorher als Häufigkeitsverteilungen, erstellt (4). Durch die stark erhöhte Datenmenge konnte die Auflösung (Volumenwahrscheinlichkeitsdichteverteilungen: Faktor 100) verbessert werden. Damit wurde zu einem die Genauigkeit der Integralbildung verbessert und zum anderen die vorhandenen Feinstrukturen innerhalb der Verteilungen sichtbar bzw. geltend gemacht (5). Bei der Erzeugung der neuen Datenbanken wurden wesentlich mehr Abstandsklassen gebildet und somit entsprechend mehr Verteilungen erzeugt. Damit sollte erreicht werden, dass bei der Berechnung der wahrscheinlichsten Zuordnung für ein zwei- oder dreideutiges NOESY-Signal möglichst jede der jeweils vorhandenen Zuordnungsmöglichkeiten durch eine geeignete bzw. repräsentative Verteilung vertreten werden kann. Bei dem bisher benutzten

	FRÜHERE DATENBANK	NEUE DATENBANKEN
Anzahl vorhandener Abstandsklassen bzw. Verteilungen	1577	3620 (Datenbank 1) 16483 (Datenbank 2) 220280 (Datenbank 3)
Datenauflösung einer Verteilung	Max. 200 Datenpunkte/ Schrittweite 0.1Å	10000 Datenpunkte (für Volumenwerteilung.) 2000 Datenpunkte (für Abstandsverteilung.)
Verteilungsart	Abstandshäufigkeitsverteilungen	Volumen/Abstands-Wahrscheinlichkeits-dichteverteilungen (Erzeugung über Summierung von Gausskurven)
Größe der Proteinstrukturdatenbasis	326 NMR Strukturen	970 Röntgenstrukturen+ 137 NMR-Strukturen
Eigenschaften der Proteinstrukturen innerhalb der jeweiligen Datenbasis	Alle Proteine wasserlöslich Keine paramagnetischen Zentren Keine Kofaktoren Keine RNA/DNA Strukturen Keine Komplexe	Keine RNA/DNA Strukturen Paarweise Sequenzidentität<25%

Abbildung 4.1: Unterschiede zwischen der alten und neuen Datenbank

Datensatz war das, aufgrund der geringeren vorhandenen Anzahl von Verteilungen, oft nicht möglich. Dies führte oft dazu, dass unterschiedlichen Zuordnungsmöglichkeiten durch die

gleiche Verteilung vertreten wurden und somit die gleichen Wahrscheinlichkeitswerte erhielten (6).
Abbildung 4.1 zeigt eine Übersicht über die wesentlichen Unterschiede zwischen der früheren Datenbank und den neuen Datenbanken *(Datenbank 1-3).*

4.1.1.2 Bildung von Abstandsklassen

Wesentliches Ziel bei der Erstellung der neuen Datenbanken war es, möglichst für alle vorkommenden Zuordnungsmöglichkeiten (=Atompaare) für NOESY-Signale eine repräsentative Abstands- oder Volumenwahrscheinlichkeitsdichteverteilung zu erhalten. Jede der erzeugten Verteilungen wurde hierbei aus einer anderen Gruppe von Atomabständen (=Abstandsklasse) generiert. Eine bestimmte Abstandklasse bezeichnet hierbei eine Gruppe von Abständen zwischen Atomen mit bestimmten definierten Eigenschaften. Hierbei sind insbesondere Eigenschaften vom Interesse, welche sich auf die Größe des räumlichen Abstandes innerhalb eines Proteins auswirken können. Es wurden folgende Kriterien bei der Auswahl von Atompaaren für die Bildung einer bestimmten Abstandsklasse angewendet:

1. Atomname innerhalb der entsprechenden Aminosäure.
2. Zugehörigkeit der Atome zu einem bestimmten Aminosäuretyp.
3. Relativer Sequenzabstand der Atome innerhalb der Proteinsequenz.
4. Reihenfolge der Atome innerhalb der Proteinsequenz.

Als Atomnamen kommen alle Wasserstoffatomnamen (nach IUPAC) in den 20 natürlich vorkommenden Aminosäuren in Frage (1) (s. Anhang A). Für den Aminosäuretyp können alle 20 natürlich vorkommenden Aminosäuren eingesetzt werden (2). Der relative Sequenzabstand ist der Differenzbetrag der Sequenzpositionen von zwei Aminosäuren innerhalb einer Proteinsequenz (3). Die Reihenfolge zweier Atome 1 und 2 bezüglich ihrer Sequenzpositionen wird im folgendem mit einem Pluszeichen versehen, wenn sich das Atom 1 näher am *n-terminalen* Ende der Proteinsequenz befindet. Im umgekehrten Fall entsprechend mit einem Minuszeichen. Bei Berücksichtigung aller genannten Auswahlkriterien für Atompaare lässt sich eine bestimmte Abstandklasse bzw. Gruppe von Atompaaren folgendermaßen definieren:

[Atomname 1, Rest 1, Atomname 2, Rest 2, i +/- ΔS]

Dabei steht der *Atomname 1* bzw. *Atomname 2* für die jeweiligen Atomnamen des Paares, *Rest 1* und *Rest 2* für die Aminosäurezugehörigkeit,+ bzw. – für die Reihenfolge innerhalb der Sequenz (s.o.) und ΔS für den relativen sequentiellen Abstand der zugehörigen Aminosäuren innerhalb einer Proteinstruktur. Der Buchstabe *i* steht hierbei für Sequenzposition der Aminosäure in der sich das zuerst genannte Atom befindet.

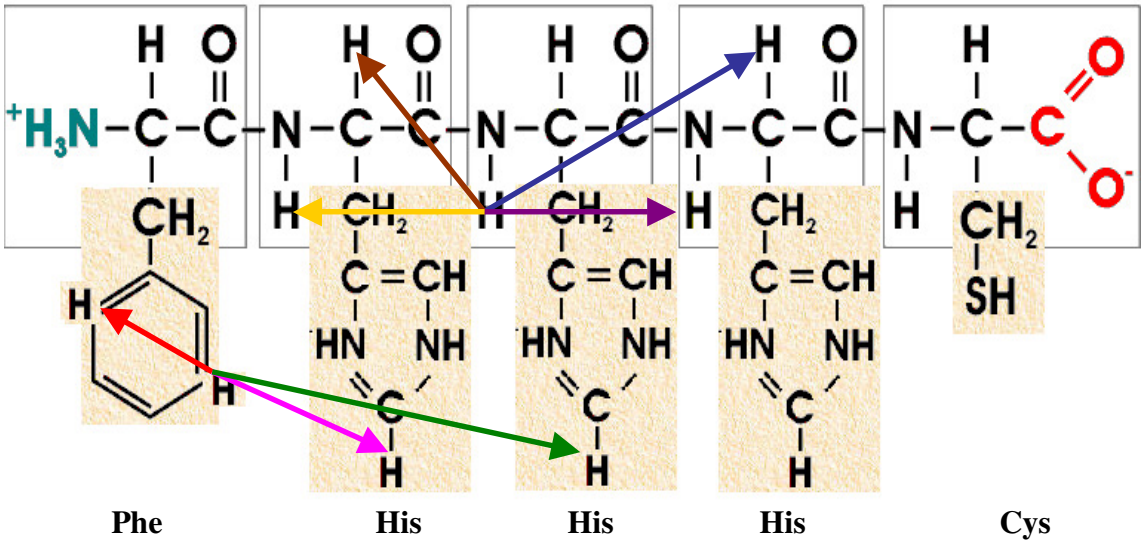


Abbildung 4.2: Beispiele von Abstandsklassen. Folgende sieben verschiedene Typen von Abstandsklassen sind anhand der hier dargestellten kurzen Peptidkette verdeutlicht:
1. [HE1, Phe, HD2, Phe, i+0] (rot), 2. [HE1, Phe, HE1,His, i+1] (rosa)
3. [HE1, Phe, HE1, His, i+2] (grün) 4. [NH, His, NH, His, i-1] (orangefarben)
5. [NH, His, NH, His, i+1] (violett) 6. [HN, His, HA, His, i+1] (blau)
7. [HN, His, HA, His, i-1] (braun)
An den Abstandsklassen 6 und 7 kann man anhand der unterschiedlichen Zahl auseinanderliegender Bindungen sehen, dass die Größe des räumlichen Abstandes bei unterschiedlicher Reihenfolge zweier vergleichbarer Atome innerhalb einer Proteinsequenz sehr unterschiedlich sein kann.

In Abbildung 4.2 sind die genannten Kriterien zur Bildung von Abstandsklassen anhand von einigen Beispielen verdeutlicht.

In der Arbeit wurden zu Versuchzwecken auch Wahrscheinlichkeitsdichteverteilungen aus Abstandsklassen erzeugt, bei denen nicht alle der oben aufgeführten Kriterien zur Bildung einer bestimmten Abstandsklasse zugleich angewendet wurden. Im ersten Fall (*Datenbank 1*) wurde entweder nicht nach dem Atomnamen und im zweiten Fall (*Datenbank 2*) nicht nach der Zugehörigkeit zu eine bestimmten Aminosäuretyp der Atome unterschieden. Das heißt, dass in den entsprechenden Wahrscheinlichkeitsdichteverteilungen Atomabstände von Atomen mit entweder unterschiedlichen Atomnamen oder unterschiedlicher Aminosäureherkunft zusammengefasst worden sind. Bei der *Datenbank 1* wurden hierbei jeweils die Abstände aus allen möglichen Wasserstoffkombinationen zweier bestimmter

Aminosäuren in eine Verteilung integriert. Es wurden insgesamt drei Datenbanken aus Wahrscheinlichkeitsdichteverteilungen unter Bildung jeweils unterschiedlicher Abstandsklassen generiert:

Datenbank 1: [Rest 1, Rest 2, i+/- ΔS]

Datenbank 2: [Atomname 1, Atomname 2, i+/- ΔS]

Datenbank 3: [Atomname 1, Rest 1, Atomname 2, Rest 2, i+/- ΔS]

Zum Vergleich sind die gebildeten Abstandsklassen für die neuen Datenbanken (s. Abb. 4.3) wie auch für die frühere Datenbank (s. Abb. 4.4) detailliert aufgeführt. Aus der Abbildung 4.3 ist zu entnehmen, dass bei allen drei neuen Datenbanken gesonderte Abstandsklassen für ΔS= 0, 1, 2, 3, 4, 5, 6, 7 und 8 erzeugt wurden. Da der mittlere räumliche Abstand von Atompaaren mit relativen Sequenzabständen bis zu etwa 12 Aminosäuren relativ stark ansteigt (s. Abb. 4.8), war zu erwarten, dass sich die daraus resultierenden Wahrscheinlichkeitsdichteverteilungen entsprechend stark voneinander unterscheiden werden.

Relativer Sequenzabstand	Atomname 1	Rest 1	Atomname 2	Rest 2	Datenbank	Anzahl vorhandener Verteilungen
0,1,2,3,4,5,6,7,8;	Zusammengefasst	1,...,20*	Zusammengefasst	1,...,20	1	**3620**
>8	Zusammengefasst	1,...,20	Zusammengefasst	1,...,20	1	
0,1,2,3,4,5,6,7,8;	1,...,42**	Zusammengefasst	1,...,42	Zusammengefasst	2	**16483**
>8	1,...,42	Zusammengefasst	1,...,42	Zusammengefasst	2	
0,1,2,3,4,5,6,7,8;	1,...,42	1,...,20	1,...,42	1,...,20	3	**220280**
>8	1,...,42	1,...,20	1,...,42	1,...,20	3	

Abbildung 4.3: Abstandsklassen der neuen Datenbanken. Die Tabelle zeigt für welche Abstandsklassen jeweils separate Wahrscheinlichkeitsdichteverteilungen erzeugt wurden. In der ersten Spalte stehen alle berücksichtigten relativen sequenziellen Abstände zweier Atome. Das " >" bedeutet, dass alle Abstände von Atompaaren die einen größeren relativen Abstand als den angegebenen Wert haben, innerhalb einer Verteilung zusammengefasst wurden. In den Spalten 2 und 4 stehen jeweils die möglichen Wasserstoffatomnamen und in den Spalten 3 und 5 alle in Betracht kommenden Aminosäuren. Allerdings sind nicht alle Kombinationen möglich, da nicht jeder Wasserstoffatomname in jeder Aminosäure vorkommt. Für Abstandsklassen mit $\Delta S=0$ (intraresiduale Abstände) sind nur Atompaarkombinationen mit Rest1=Rest2 und Atomname1 # Atomname2 sinnvoll.

**Alle 20 natürlich vorkommenden Aminosäuren*

***Alle in den 20 natürlichen Aminosäuren vorkommenden Wasserstoffatomnamen (nach IUPAC)*

Sequenzabstand	Atomname 1	Rest 1	Atomname 2	Rest 2
0	HN	1,..,20	1,..,41 außer HN(=42)	1,..,20
0	1,..,41 außer HN(=42)	Zusammengefasst	HN	Zusammengefasst
1	1,...,42	Zusammengefasst Keine Ringprotonen	HN	Zusammengefasst
1	1,...,42	Zusammengefasst Keine Ringprotonen	HA	Zusammengefasst
1	1,..,41 außer HN(=42)	Zusammengefasst	Ringprotonen	Aromatische Reste
1	Ringprotonen	Aromatische Reste	1,...,42	Zusammengefasst
2	1,...,42	Zusammengefasst Keine Ringprotonen	HN	Zusammengefasst
2	1,...,42	Zusammengefasst Keine Ringprotonen	HA	Zusammengefasst
2	1,..,41 außer HA(=42)	Zusammengefasst	Ringprotonen	Aromatische Reste
2	Ringprotonen	Aromatische Reste	1,...,42	Zusammengefasst
3	1,...,42	Zusammengefasst Keine Ringprotonen	HN	Zusammengefasst
3	1,...,42	Zusammengefasst Keine Ringprotonen	HA	Zusammengefasst
3	1,..,41 außer HA(=42)	Zusammengefasst	Ringprotonen	Aromatische Reste
3	Ringprotonen	Aromatische Reste	1,...,42	Zusammengefasst
4	1,...,42	Zusammengefasst Keine Ringprotonen	HN	Zusammengefasst
4	1,...,42	Zusammengefasst Keine Ringprotonen	HA	Zusammengefasst
4	1,..,41 außer HA(=42)	Zusammengefasst	Ringprotonen	Aromatische Reste
4	Ringprotonen	Aromatische Reste	1,...,42	Zusammengefasst
>4	1,...,42	Zusammengefasst Keine Ringprotonen	HN	Zusammengefasst
>4	1,...,42	Zusammengefasst Keine Ringprotonen	HA	Zusammengefasst
>4	1,..,41 außer HA(=42)	Zusammengefasst	Ringprotonen	Aromatische Reste
>4	Ringprotonen	Aromatische Rest	1,...,42	Zusammengefasst

Abbildung 4.4: Abstandsklassen der früheren Datenbank

4.1.2 Beispiele für Wahrscheinlichkeitsdichte-verteilungen

Hier soll an einigen Beispielen ein Eindruck von den in dieser Arbeit erzeugten Abstands- und Volumenwahrscheinlichkeitsdichteverteilungen vermittelt werden. Dabei wird gezeigt, wie sich unterschiedliche angewandte Kriterien bei der Bildung von Abstandsklassen auf den Kurvenverlauf einer bestimmten Verteilung auswirken können.

4.1.2.1 Abstands - und Volumenwahrscheinlichkeitsdichte-verteilungen

Der wesentliche Unterschied bei der Erzeugung einer Volumenwahrscheinlichkeitsdichte-verteilung gegenüber einer Abstandswahrscheinlichkeitsdichteverteilung war, dass die für eine bestimmte Abstandsklasse gefundenen Abstände r [Å] vorher mit der Beziehung $V=(1/r^6)* Å^9$ in ein Signalvolumen *V* umgewandelt wurden. Aus den resultierenden Werten wurde daraufhin die entsprechende Verteilung berechnet.

Die Diagramme *A* und *B* in Abbildung 4.5 zeigen für die Abstandsklasse [HE3, Trp, HN, Gly, (i+2)] die resultierenden Abstands- bzw. Volumenwahrscheinlichkeitsdichteverteilungen und die Diagramme *C* und *D* die dazugehörigen Häufigkeitsverteilungen. Es fällt auf, dass, neben dem relativ stark unterschiedlichen Kurvenverlauf zwischen Abstands- und Dichteverteilungen, innerhalb der Volumenwahrscheinlichkeitsdichteverteilung wesentlich weniger Feinstrukturen sichtbar sind. Der Grund dafür ist, dass sich die Werte (Volumina /Abstände), innerhalb der gebildeten Volumen bzw. Abstandsintervalle unterschiedlich stark verteilen. Bei den Volumenwahrscheinlichkeitsdichteverteilungen verteilen sich die vorhandenen Werte, trotz der großen Anzahl gebildeter Intervalle (10 000), sehr ungleichmäßig auf die vorhandenen Intervalle auf. Wie man aus der Volumenhäufigkeitsverteilung entnehmen kann (Abb. 4.5 *D*), decken die ersten wenigen (von 10000) Volumenintervalle bereits relativ große Abstandsbereiche ab. Das führt dazu, dass die innerhalb der Abstandswahrscheinlichkeitsdichteverteilungen sichtbaren Feinstrukturen innerhalb der entsprechenden Volumenwahrscheinlichkeitsdichteverteilungen oft unsichtbar werden.

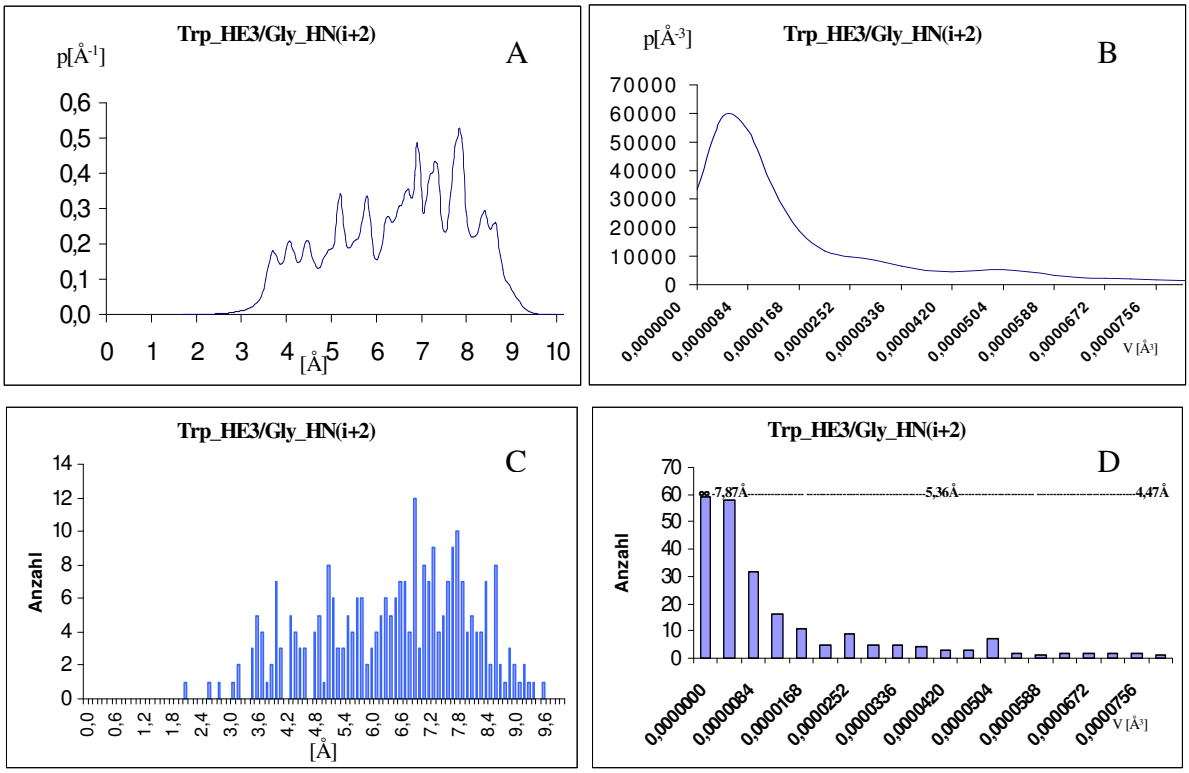


Abbildung 4.5: Vergleich zwischen Abstands- und Volumenwahrscheinlichkeitsdichteverteilungen. Die oberen beiden Grafiken zeigen jeweils für die Abstandsklasse [HE3, Trp, HN, Gly, (i+2)] die zugehörige Abstands- (A) bzw. Volumenwahrscheinlichkeitsdichteverteilung (B). In den unteren beiden Grafiken (C, D) sind die entsprechenden Häufigkeitsverteilungen abgebildet. Die Datenauflösung bzw. Intervallbreite beträgt bei der gezeigten Abstandswahrscheinlichkeitsdichteverteilung und Abstandshäufigkeitsverteilung jeweils 0.1 Å. Für die Volumenwahrscheinlichkeitsdichteverteilung und Volumenwahrscheinlichkeitshäufigkeitsverteilung beträgt der entsprechende Wert 0.0000042 $Å^3$.

4.1.2.2 Identifikation von Sekundärstrukturen

Hier soll gezeigt werden, dass sich innerhalb in dieser Arbeit erzeugten Abstandswahrscheinlichkeitsdichteverteilungen Sekundärstrukturen identifizieren lassen. In Abbildung 4.6 sind Abstandswahrscheinlichkeitsdichteverteilungen für *HA* und *HN* Atome für relative Sequenzabstände von je 1,2,3 und 4 Aminosäuren dargestellt. Die Sekundärstrukturen, die für ein bestimmtes Maximum im Kurvenverlauf verantwortlich sind, sind in der jeweiligen Grafik angegeben.

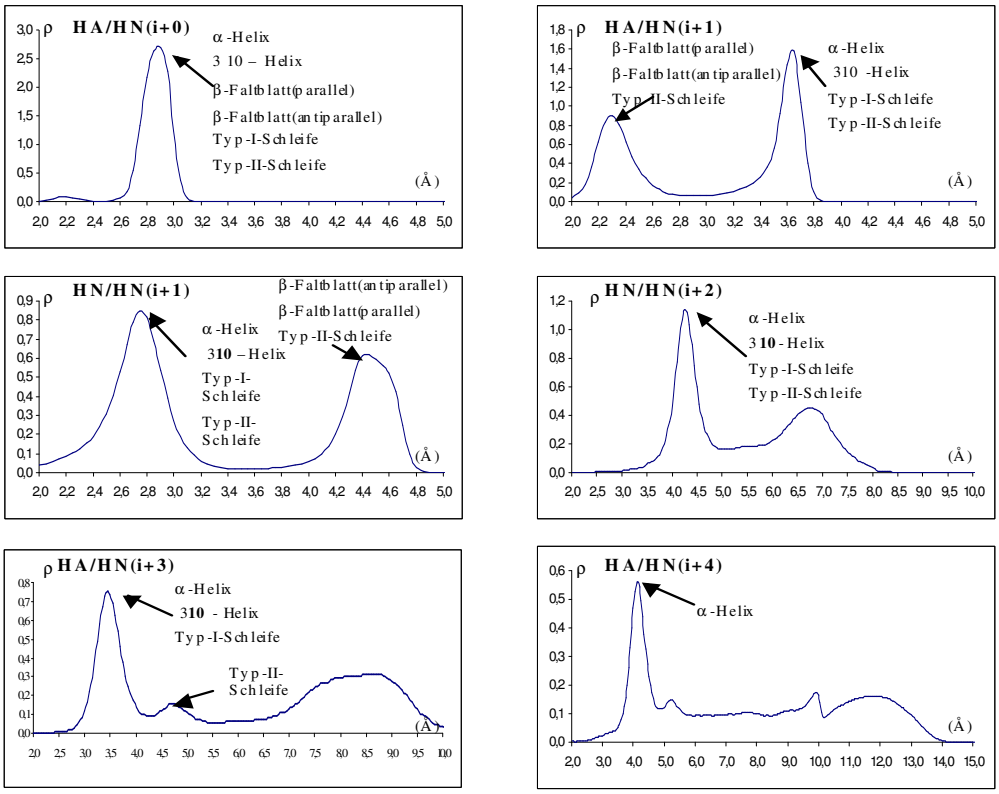


Abbildung 4.6: Identifizierung von Sekundärstrukturen. Die Abbildung zeigt einige Abstandswahrscheinlichkeitsdichteverteilungen von *HA* und *HN* Atompaaren für jeweils unterschiedliche relative Sequenzabstände. Charakteristische Atomabstände, wie sie in verschiednen Sekundärstrukturen (s. Anhang *C*) vorkommen, machen sich als Maxima (Pfeile) in den Kurvenverläufen bemerkbar.

4.1.2.3 Wahrscheinlichkeitsdichteverteilungen unterschiedlicher Abstandsklassen

Hier soll an einigen Beispielen ein Eindruck davon vermittelt werden, wie sich die Bildung unterschiedlicher Abstandsklasse auf den Verlauf einer Wahrscheinlichkeitsdichteverteilung auswirken kann. In der Abbildung 4.7 sind die Verläufe von einigen Volumenwahrscheinlichkeitsdichteverteilungen für folgende Abstandklassen dargestellt:

1. Spalte: [Gly, Pro, i + (1,2,. 5)]

2. Spalte: [HN, HD2, i + (1,2,..5)]

3. Spalte: [HN, Pro, HD2, Gly, i + (1,2,..5)]

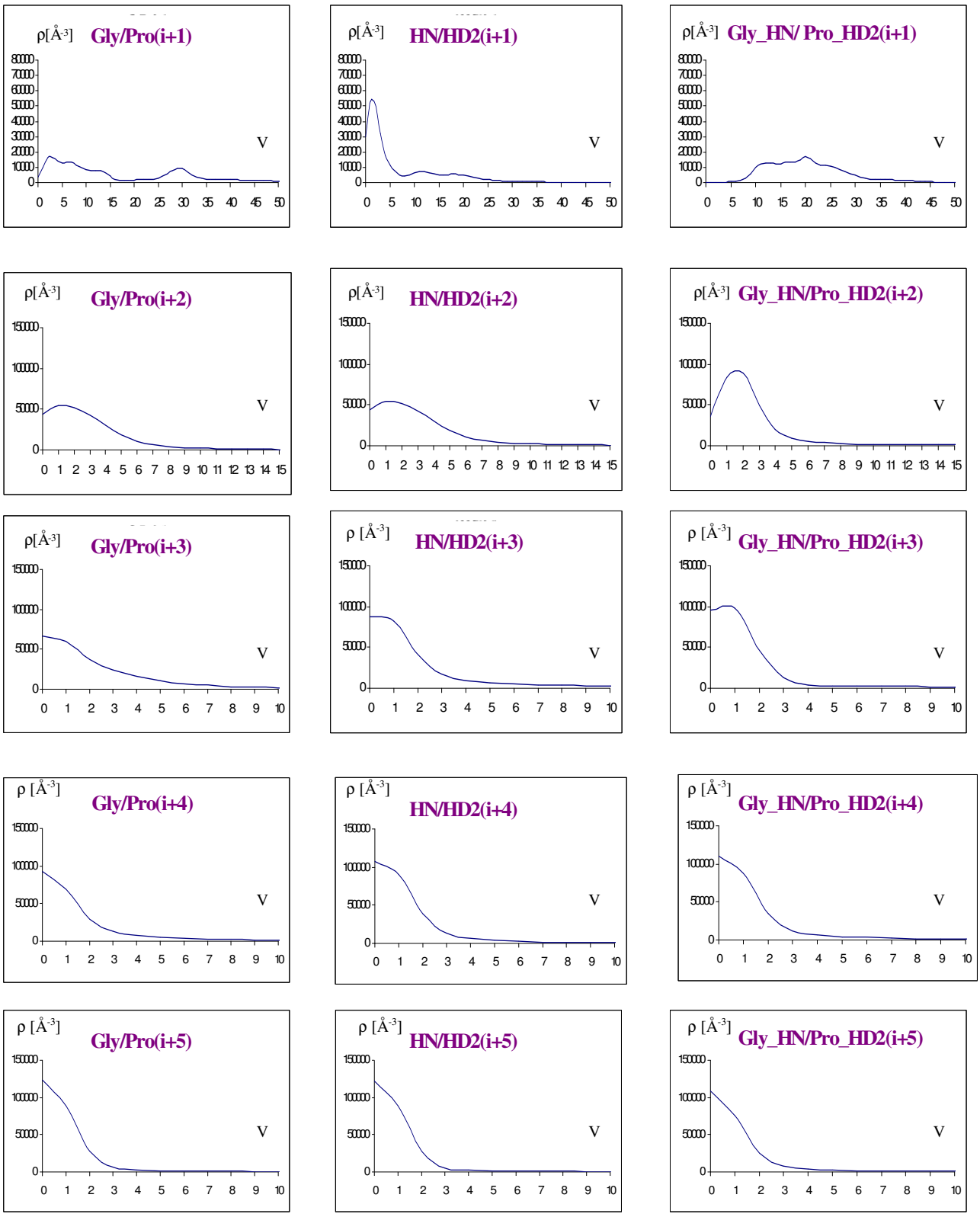


Abbildung 4.7: Volumenwahrscheinlichkeitsdichteverteilungen für unterschiedliche Abstandsklassen. Die auf der Abszisse angegebenen Volumenwerte (V) entsprechen einem Vielfachen von 0.0000042 Å^3.

Wie man sieht, unterscheiden sich die Kurvenverläufe der verschiedenen Typen von Volumenwahrscheinlichkeitsdichteverteilungen für kleine Werte von ΔS deutlich voneinander. Mit Zuname des Sequenzabstandes gleichen sich die Kurvenprofile allerdings immer weiter aneinander an. Abbildung 4.8 zeigt den Zusammenhang zwischen der Größe des mittleren räumlichen Abstandes für die drei oben genannten Abstandklassen in Abhängigkeit vom jeweiligen relativen Sequenzabstand des Atompaares.

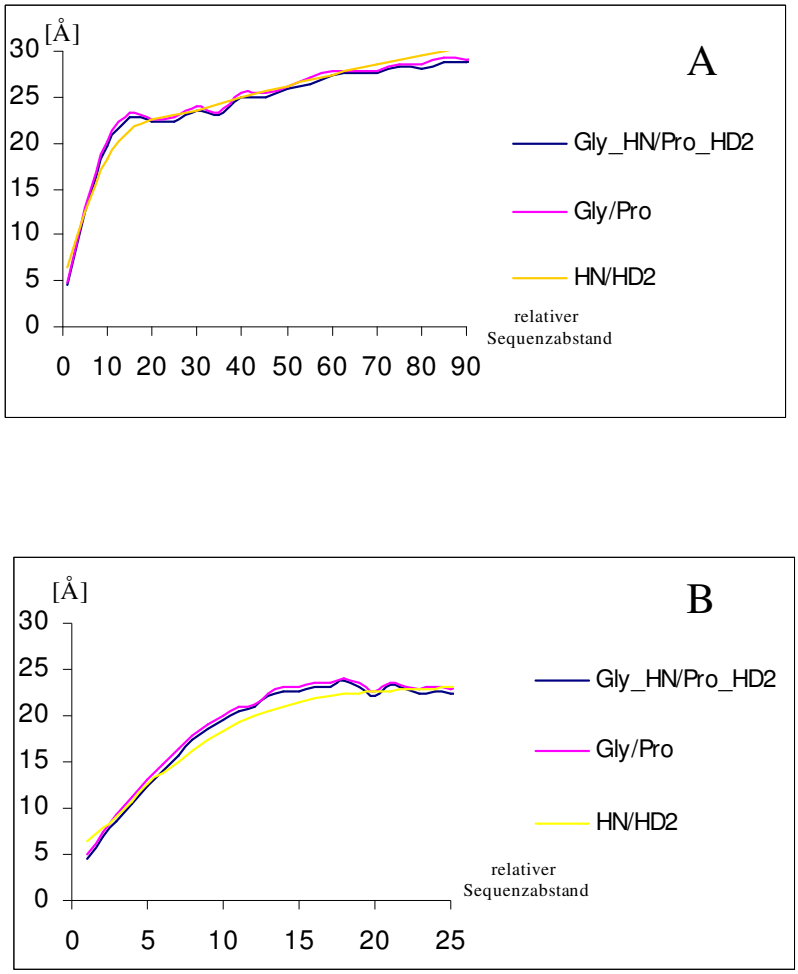


Abbildung 4.8: Mittlere räumliche Abstände Die Grafiken zeigen die Abhängigkeit der mittleren räumlichen Abstände vom relativen Sequenzabstand der Atome. Hier gezeigt für die Abstandsklassen [HN, Gly, HD2, Pro, i+ΔS] (schwarz), [Gly, Pro, i+ΔS] (rosa) und [HN, HD2, i+ΔS] (gelb) mit ΔS={1,2,...90}. In der Grafik A ist im Abstand von fünf Aminosäuren ein Wert aufgetragen. In Grafik B beträgt die Schrittweite je eine Aminosäure.

Man kann sehen, dass der Verlauf der Kurven für relative Sequenzabstände, bis etwa 10 Aminosäuren, relativ steil linear ansteigt und sich dann schnell abflacht. Die Grafiken (*a-d*) in Abbildung 4.9 zeigen vier Abstandswahrscheinlichkeitsdichteverteilungen der Abstandklassen:

1. [Lys, Lys, i+0]
2. [HN, Lys, HB, Lys, i+0]
3. [HN, Lys, HG3, Lys, i+0]
4. [HN, Lys, HD3, Lys, i+0]

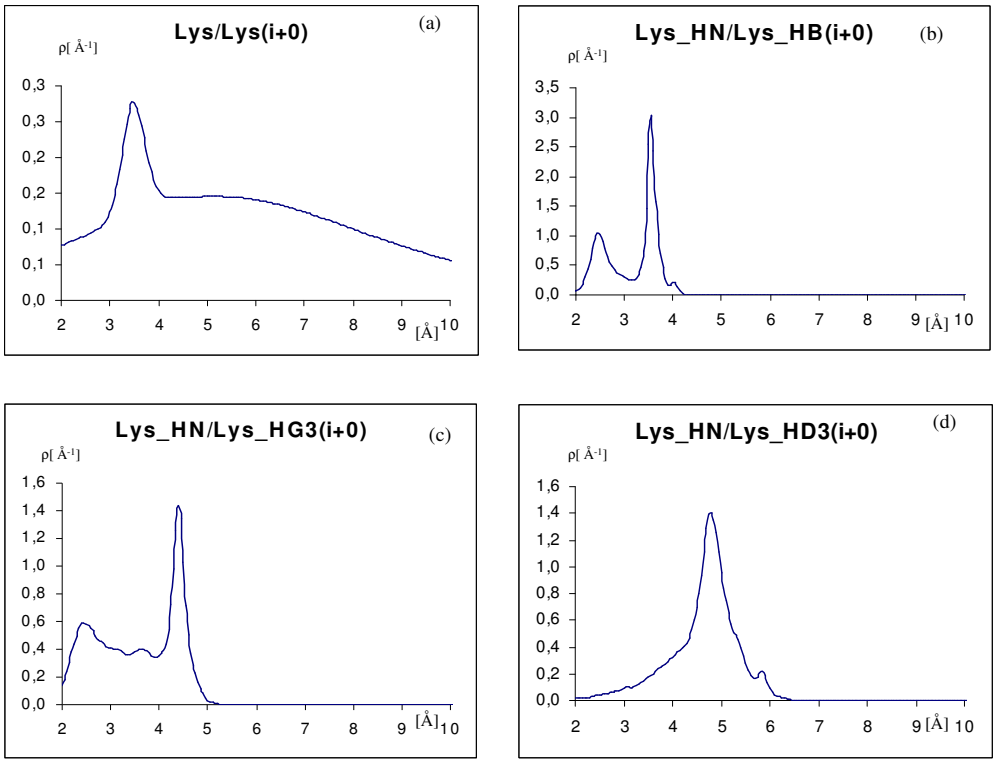


Abbildung 4.9: Abstandswahrscheinlichkeitsdichteverteilungen intraresidualer Atome. Die Grafiken (a-d) zeigen jeweils Abstandswahrscheinlichkeitsdichteverteilungen für die Abstandsklassen [Lys,Lys,i+0], [HN,Lys,HB,Lys,i+0], [HN,Lys,HG3,Lys,i+0] und [HN,Lys,HD3,Lys, i+0].

Wie hier zu sehen ist, führte hier die Berücksichtigung des Atomnamens bei der Abstandsklassenbildung zu deutlich unterschiedlichen Kurvenverläufen (Abb. 4.9 *c-d*). In den Grafiken *a-e* der Abbildung 4.10 kann man erkennen, wie die Nichtberücksichtung der Aminosäurezugehörigkeit der Atome bei der Abstandklassenbildung zu beinahe annäherndem Verschwinden der vorhandenen Maxima führen kann.

In der Abbildung 4.10 sind 5 Abstandswahrscheinlichkeitsdichteverteilungen folgender Abstandklassen zu sehen:

1. [HE1, HB, i+0]
2. [HE1, Phe, HB2, Phe, i+0]
3. [HE1, Tyr, HB2, Tyr, i+0]
4. [HE1, His, HB2, His, i+0]
5. [HE1, Trp, HB2, Trp, i+0]

In den Verteilungen der Abstandklassen 2-5 (Grafiken *b-e*) sind zwei deutlich getrennte Maxima zu sehen.

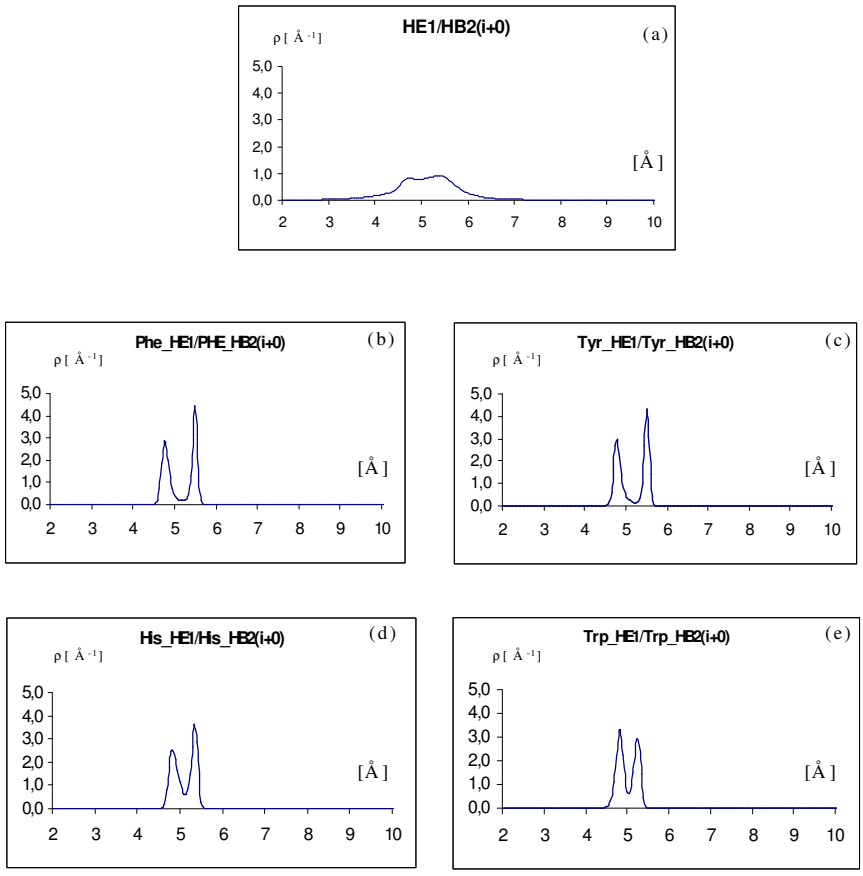


Abbildung 4.10: Abstandswahrscheinlichkeitsdichteverteilungen intraresidualer Atomabstände. Die Grafiken (a-e) zeigen jeweils Abstandswahrscheinlichkeitsdichteverteilungen für die Abstandsklassen [HE1,HB2,i+0], [HE1,Phe,HB2,Phe,i+0], [HE1,Tyr,HB2,Tyr,i+0], [HE1,His,HB2, His,i+0], [HE1, Trp,HB2, Trp,i+0].

4.1.2.4 Die Bedeutung der Datenauflösung

Hier soll anhand eines Beispiels gezeigt werden, wie sich die Datenauflösung auf den Verlauf einer Wahrscheinlichkeitsdichteverteilung auswirken kann.

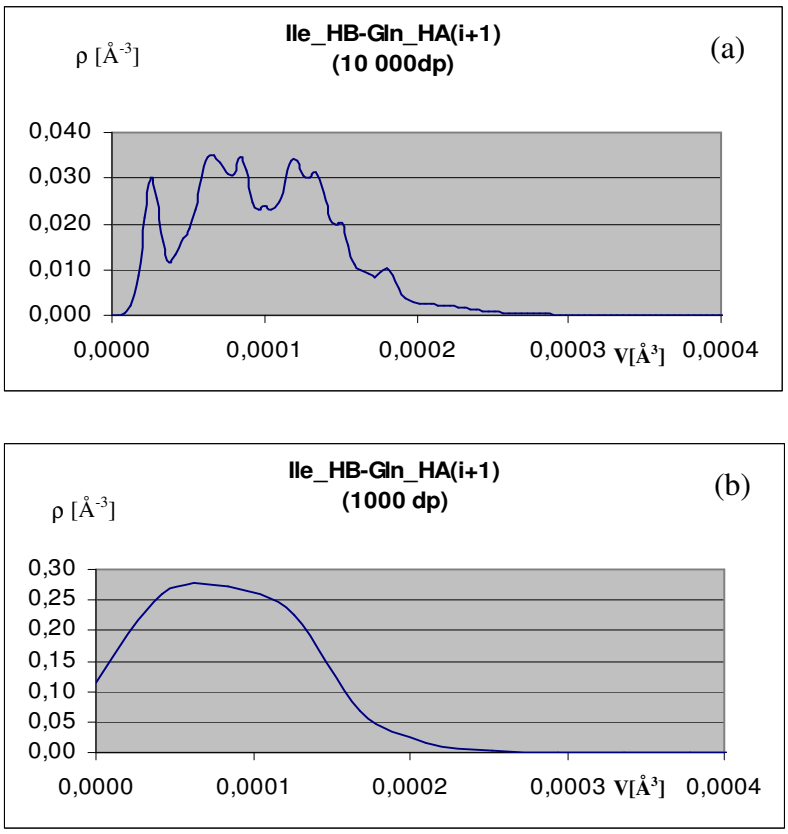


Abbildung 4.11: Verlust des Kurvenprofils durch Herabsetzung der Datenauflösung.

Die dargestellten Diagramme zeigen die Volumenwahrscheinlichkeitsdichteverteilung für die sequentiell benachbarten Atome *HB* in der Aminosäure Isoleucin, *HA* und Glutamat. Die Kurven in den Diagrammen haben jeweils eine Datenauflösung von 10000 (Abb.4.11 *a*) bzw. 1000 (Abb. 4.11 *b*) Datenpunkten. Die Reduktion der Datenauflösung von 10000 auf 1000 Datenpunkte wurde durch das Zusammenfassen von jeweils 10 hintereinanderliegenden Werten zu einem Mittelwert erzielt. In Abbildung 4.11 *b* kommt es durch die Reduzierung der Datenauflösung zu einem starken Verlust des Kurvenprofils. Daraus lässt sich schließen, dass die hier gewählte hohe Auflösung der Verteilungen wichtig für die Sichtbarmachung vorhandener Feinstrukturen ist.

4.2.Überprüfung der Zuordnungsqualität unter Benutzung der neuen Datenbanken

4.2.1 Einfluss des Suchradius und der Toleranz der chemischen Verschiebung auf die Zuordnungs-möglichkeiten

Hier soll zunächst untersucht werden, wie sich die eingestellte Toleranz der chemischen Verschiebung und der Suchradius auf die Eigenschaften, der sich für die NOESY-Signale ergebenen, Zuordnungsmöglichkeiten auswirkt. Diese haben nämlich, wie sich noch zeigen wird, einen starken Einfluss auf die Zuordnungsqualität. Im Zentrum der Betrachtung stehen hierbei vor allem zwei- und dreideutig zugeordnete NOESY-Signale, deren automatische Zuordnung im Rahmen dieser Arbeit verbessert werden soll. Die hier erstellte Analyse war eine wichtige Grundlage für die Wahl der Toleranz der chemischen Verschiebung, um bei verschiedenen hier getesteten Suchradien (s. Kap. 4.2.2) genügend zwei- und dreideutige NOESY-Signale für eine statische Auswertung zu Verfügung zu haben. Außerdem soll auch ein Eindruck vermittelt werden, im welchem Umfang mehrdeutige NOESY-Signale bei unterschiedlichen gewählten Werten der genannten Parameter auftreten.

4.2.1.1 Anzahl mehrdeutiger NOESY-Signale

Wie bereits im Kapitel 2.5.2 erwähnt, sind der Suchradius und die Toleranz der chemischen Verschiebung diejenigen Eingangsparameter im Programm *KNOWNOE*, über die der Benutzer die Anzahl der Zuordnungsmöglichkeiten für ein NOESY-Signal einschränken kann. Auf die sinnvolle Wahl dieser Parameter ist bereits im Kapitel 2.5.4 eingegangen worden. Im folgendem soll anhand der simulierten 2D-NOESY-NMR Spektren der Proteine *Tm*CSP und HPr gezeigt werden, wie sich die Wahl unterschiedlicher Werte der oben genannten Parameter auf die Verteilung eindeutiger, zweideutiger, dreideutiger und mehr als dreideutiger Zuordnungen auswirkt (Abb. 4.12 *Tm*CSP / 4.13 HPr). Es ist zu erkennen, dass eine Vergrößerung des Toleranzwertes der chemischen Verschiebung bei konstanten Suchradius zu einer Verschiebung der Anteile zu mehrdeutigen NOESY-Signalen bzw. zu einer Abnahme des Anteils von eindeutig zugeordneten NOESY-Signalen führt. Eine Vergrößerung des Suchradius bei konstantem eingestelltem Toleranzwert der chemischen Verschiebungen hat denselben Effekt.

Bei genauerer Betrachtung der Diagramme fällt auf, dass bei einem sehr kleinen Suchradius (0,6 nm) die Summe aus eindeutig, zweideutig, dreideutig und mehr als dreideutigen Signalen kleiner ist, als die Anzahl aller vorhandenen NOESY-Signale im Spektrum. So ergeben sich bei einem Suchradius von 0,6 nm und einem Toleranzwertwert der chemischen Verschiebung von 0,01 ppm, anstatt 2736 Signale (für *Tm*CSP), 2690 Signale bzw., anstatt 3178 Signale (für HPr *),* 3136 Signale. Der Grund dafür ist, dass einige NOESY-Signale keine Zuordnung erhalten haben.

Aufgrund chemischer Verschiebungen erhalten zunächst alle vorhanden NOESY-Signale mindestens eine Zuordnung. Dies ist auch der Fall, wenn der Toleranzwert der chemischen Verschiebung auf Null gesetzt wird. Das liegt daran, dass die Informationen über die chemischen Verschiebungen der Atome aus den eben hier simulierten NMR-Spektren kommen und sich somit jedes Signal eindeutig zuordnen lässt. In diesem speziellen Fall ist somit der Toleranzwert der chemischen Verschiebung nicht wirklich nötig. Allerdings würde man keine mehrdeutigen Signale erhalten, die aber für die weiteren Tests benötigt werden. Wenn einige Signale keine Zuordnung erhalten haben liegt das daran, dass die anfangs erstellten Zuordnungen wieder entfernt wurden, da diese einen größeren Abstand innerhalb der Modellstruktur haben als der eingestellte Suchradius. Dies dürfte aber bei dem eingestellten Suchradius von 6Å, im Falle der richtigen Zuordnungsmöglichkeit, welche sich ja immer unter den vorhandenen Zuordnungsmöglichkeiten befindet (s.o.), nicht passieren, da in den hier simulierten NOESY Spektren nur Signale zwischen Atompaaren generiert worden sind, welche maximal 0,5nm auseinanderliegen.

Der Grund für dieses Phänomen ist, dass eines oder beide Atome der richtigen Zuordnungsmöglichkeit jeweils zu einer Methyl/engruppe gehören. Da allerdings in einem NOESY Spektrum Atome von Methylgruppen nicht unterschieden werden können, fasst das Programm *KNOWNOE* diese in der vorhandenen Modellstruktur, durch geometrische Mittelung, zu einem *Pseudoatom* zusammen, welches dann im weiteren für die Abstandsbestimmung verwendet wird.

Das so erzeugte *Pseudoatom* kann dabei einen größeren Abstand (größer als Suchradius) zu dem anderen Atom innerhalb der gegebenen Zuordnungsmöglichkeit einnehmen, als das vorher einzelne Atom der Methyl/engruppe.

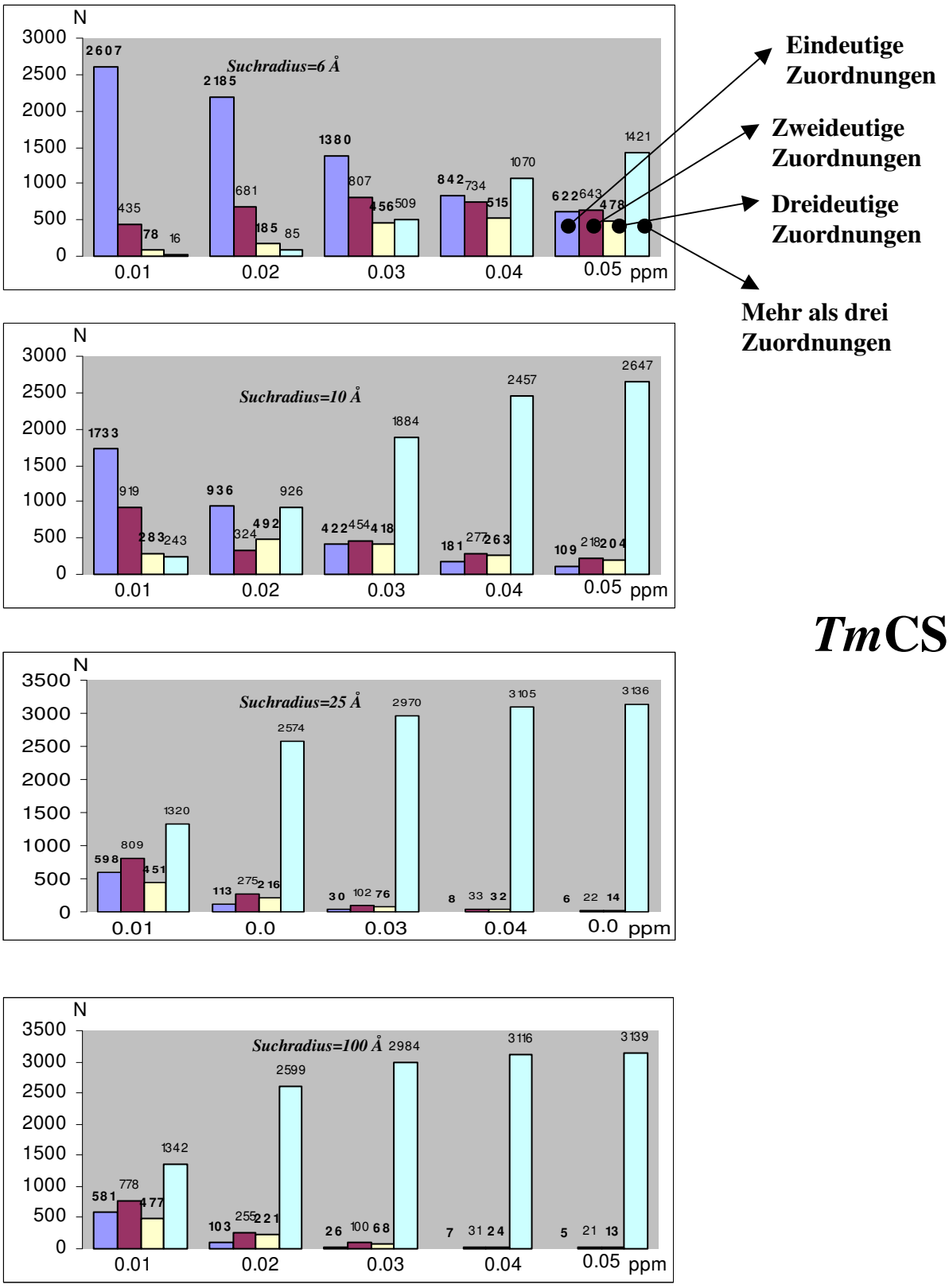
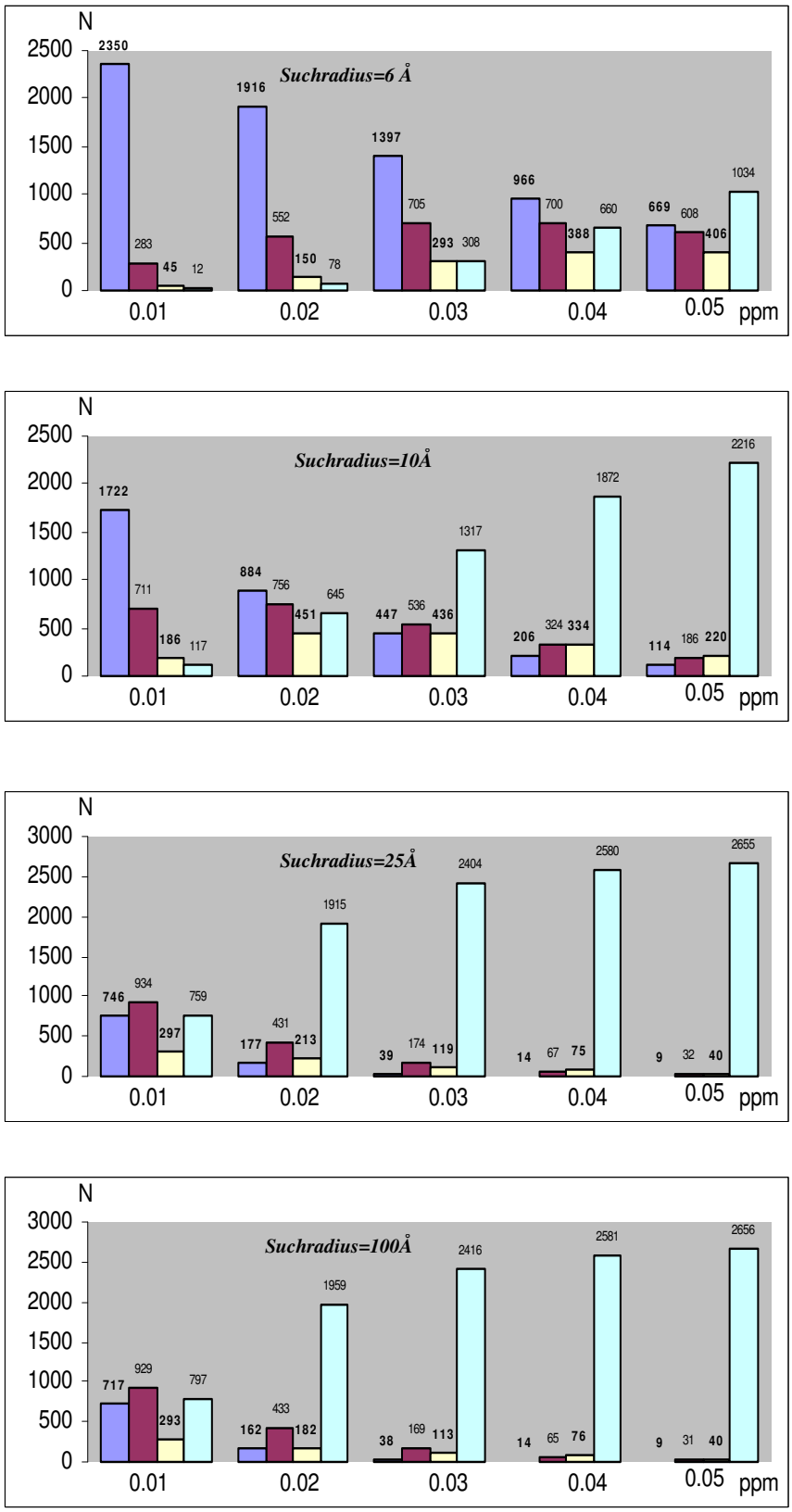


Abbildung 4.12: Verteilung von eindeutig- und mehrdeutig zugeordneten NOESY -Signalen. Hier gezeigt für das simulierte 2D-NOESY-NMR Spektrum vom Protein *Tm*CSP für unterschiedliche Kombinationen von der eingestellten Toleranz der chemischen Verschiebung und dem Suchradius. Die Gesamtzahl aller NOESY -Signale beträgt 2736.

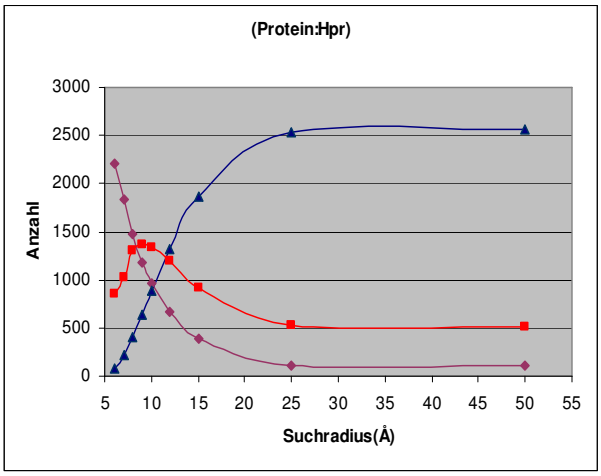


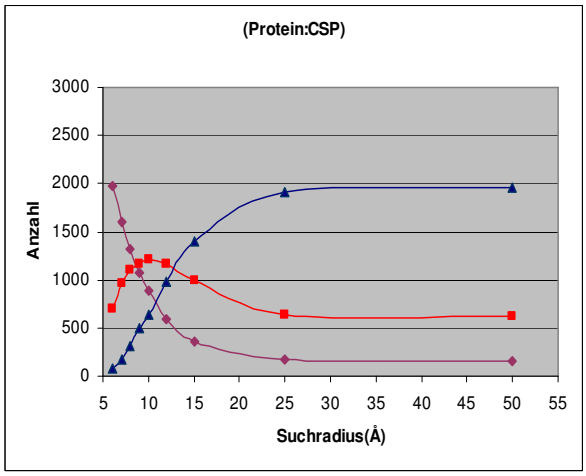
HPr

Abbildung 4.13: Verteilung von eindeutig und mehrdeutig zugeordneten NOESY-Signalen. Hier gezeigt für das simulierte 2D-NOESY-NMR Spektrum vom Protein HPr für unterschiedliche Kombinationen von der eingestellten Toleranz der chemischen Verschiebung und dem Suchradius. Die Gesamtzahl aller NOESY-Signale beträgt 3178.

4.2.1.2 Einfluss des Suchradius auf die Eigenschaften der Zuordnungen

Hier sollen unter verschiedenen eingestellten Suchradien die sich verändernden Eigenschaften der in den simulierten 2D-NOESY-NMR Spektren gefundenen Zuordnungen genauer untersucht werden. Die fest eingestellte Toleranz der chemischen Verschiebung beträgt hier 0,015 ppm (s. Kap. 3.6.2). In der Abbildung 4.14 sind die Mengen von eindeutigen, zweideutigen, dreideutigen und mehr als dreideutigen Signalen für unterschiedliche eingestellte Werte für den Suchradius angegeben. Die genauen Werte sind in den darunter liegenden Tabellen eingetragen:





Suchradius (Å)	Eindeutig	Zwei/Dreideutig	Mehr als dreideutig
6	2210	851	75
7	1834	1025	216
8	1471	1304	403
9	1175	1362	641
10	963	1334	881
12	665	1191	1322
15	389	921	1868
25	115	525	2538
50	105	508	2565

Suchradius (Å)	Eindeutig	Zwei/Dreideutig	Mehr als dreideutig
6	1976	702	78
7	1597	968	168
8	1319	1107	309
9	1075	1164	497
10	884	1207	645
12	590	1166	980
15	350	993	1393
25	177	644	1915
50	162	615	1959

Abbildung 4.14: Verteilung ein- und mehrdeutiger NOESY-Signale im Abhängigkeit vom eingestellten Suchradius. Die violette Kurve steht für die eindeutigen Signale, die rote Kurve für die zwei und dreideutigen NOESY-Signale und die blaue Kurve für NOESY-Signale mit jeweils mehr als drei erhaltenen Zuordnungen. Der eingestellte Wert für die Toleranz der chemischen Verschiebung betrug 0.015 ppm.

Wie zu erkennen ist, fällt die Anzahl der eindeutig zugeordneten Signale mit zu nehmenden Suchradius (bis etwa 25 Å) stark ab, wohingegen die Anzahl der mehr als dreideutigen NOESY-Signale sich genau umgekehrt verhält. Die Anzahl der zwei- und dreideutig

zugeordneten NOESY-Signale erreicht ihr Maximum bei einem Suchradius von etwa 10 Å und wird dann wieder kleiner.

Für die Abstandsbestimmung sind hauptsächlich solche NOESY-Signale interessant, deren Signalvolumen zum Großteil von *einem* Atompaar bestimmt wird. Abbildung 4.15 zeigt, in Abhängigkeit vom jeweils eingestellten Suchradius, den prozentualen Anteil jener zwei- und dreideutig zugeordneten NOESY-Signale, deren Signalvolumen zu mehr als 90% vom einem bestimmten Atompaar generiert wird, falls die richtige Ausgangsstruktur zur Berechnung benutzt wurde.

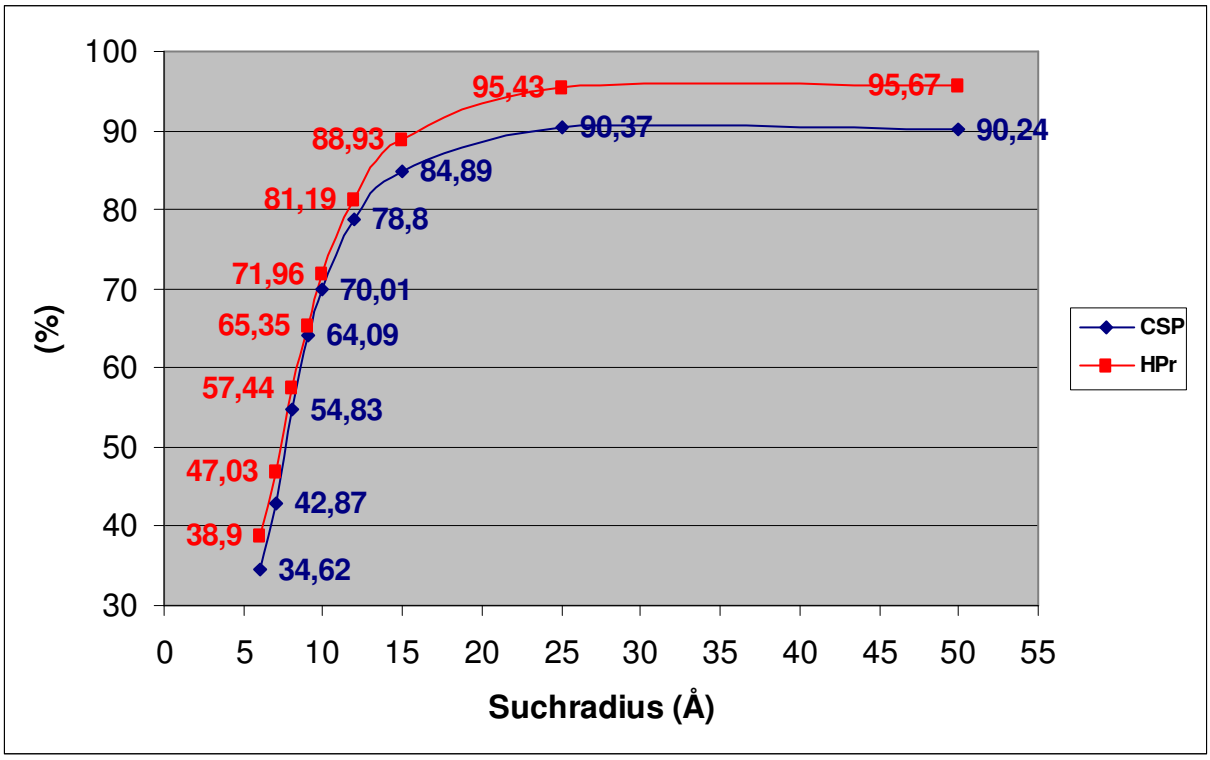


Abbildung 4.15: Anteil wichtiger NOESY-Signale. Die Grafik zeigt den prozentualen Anteil der zwei und dreideutigen NOESY-Signale, in Abhängigkeit vom Suchradius, deren Signalvolumen von einem bestimmten Atompaar zu mindestens 90% erklärt wird. Die Werte sind für die simulierten 2D-NOESY-NMR Spektren der Proteine *HPr (*rot) und *TmCSP* (blau) aufgeführt.

Es zeigt sich, dass mit kleiner werdenden eingestellten Suchradius der Anteil dieser Signale stark abnimmt. Der Grund hierfür liegt darin, dass sich die Abstände der Zuordnungs-möglichkeiten für ein NOESY-Signal innerhalb der Struktur tendenziell immer weniger voneinander unterscheiden. Somit kommt es immer öfter vor, dass keine der jeweiligen vorhandenen Zuordnungsmöglichkeiten einen Großteil des Gesamtvolumens des betreffenden NOESY-Signals erklärt.

4.2.2 Qualität der Signalzuordnungen

Hier soll vor allem auf die Verbesserung der Zuordnungsqualität von zwei- und dreideutigen NOESY-Signalen unter Einsatz der neuen Wahrscheinlichkeitsdichteverteilungen eingegangen werden.

Kernziel des hier angewandten statistischen Verfahrens zur Zuordnung von zwei- und dreideutigen NOESY-Signalen ist es, möglichst viele Signale zu identifizieren, deren Volumen von einem Atompaar zu mindestens 90% erklärt wird und dieses entsprechend zuzuordnen. Diese Zuordnungen bilden die Grundlage für eine exakte Abstandsbestimmung. Je mehr solche Zuordnungen vorhanden sind umso mehr Abstandsbeschränkungen können in die Strukturrechnung fließen. Dies führt wiederum zu einer schnelleren und genaueren Strukturbestimmung. Deshalb war eines der Ziele durch Einsatz der neuen Datenbanken die Anzahl der Zuordnungen zu erhöhen.

Da es sich bei dem hier angewandten Verfahren um eine statistische Methode handelt, kann nicht für jedes in Frage stehende NOESY-Signal die optimale Lösung gefunden werden. Folgende ungünstige Fälle können eintreten:

1. Dem in Frage stehenden NOESY-Signal wird ein Atompaar zugewiesen das nicht in Wirklichkeit den Großteil des vorhandenen Signalvolumens erklärt (hier auch als falsche Zuordnung bezeichnet).

2. Es erfolgt eine Zuordnung obwohl es kein Atompaar innerhalb der Struktur gibt, welches das betreffende NOESY-Signal allein zu mindestens 90 % erklärt (hier auch als unerwünschte Zuordnung bezeichnet).

In beiden Fällen kommt es aufgrund eines zu groß angenommenen Signalvolumens zur Berechnung eines zu kleinen Abstandes für das jeweils zugeordnete Atompaar. Deshalb war ein weiteres Ziel, durch Benutzung der neuen Datenbanken, die Anteile dieser Zuordnungen zu minimieren.

4.2.2.1 Gesamtzunahme von Zuordnungen

Hier soll gezeigt werden, inwieweit durch Benutzung der neuen Datenbanken die Gesamtzahl der zugeordneten zwei- und dreideutigen NOESY-Signalen gesteigert werden konnte.

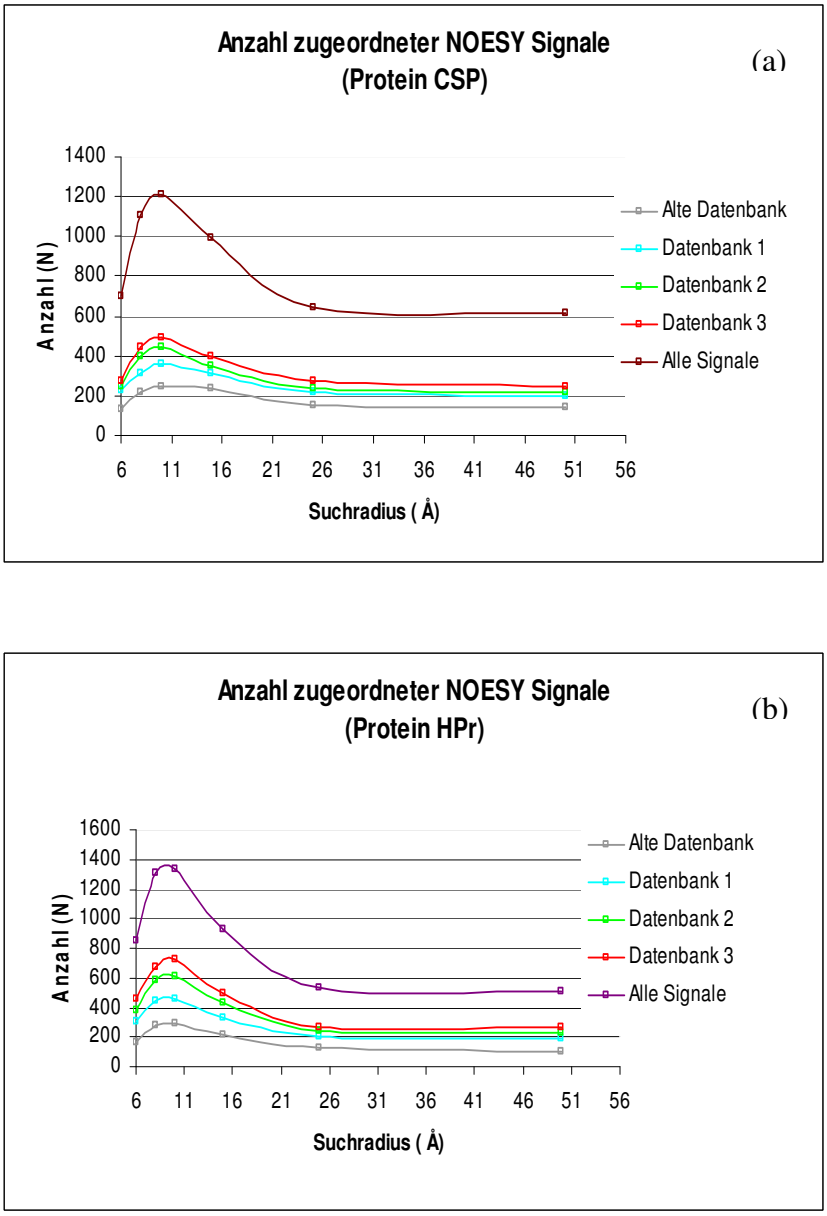


Abbildung 4.16 : Anzahl zugeordneter NOESY-Signale. Die Grafiken zeigen die Anzahl derjenigen zwei- und dreideutigen NOESY-Signale, die mit einer Wahrscheinlichkeit von 98% zugeordnet werden konnten. Die Werte sind für unterschiedliche eingestellte Suchradien und bei Benutzung verschiedener Datenbanken aufgeführt. Zum Vergleich ist auch die Anzahl aller vorhandenen zwei und dreideutigen NOESY-Signale aufgeführt (violette Kurve). Die obere Grafik (a) zeigt die Werte für das Spektrum vom Protein CSP bzw. die untere (b) für das Protein HPr.

Die Grafiken in Abbildung 4.16 zeigen die absolute Anzahl, während die Grafiken in Abbildung 4.17 den prozentualen Anteil derjenigen zwei- und dreideutig zugeordneten

NOESY-Signale zeigt, bei denen für eine der gegebenen Zuordnungsmöglichkeiten eine Wahrscheinlichkeit von mindestens 98% berechnet werden konnte.

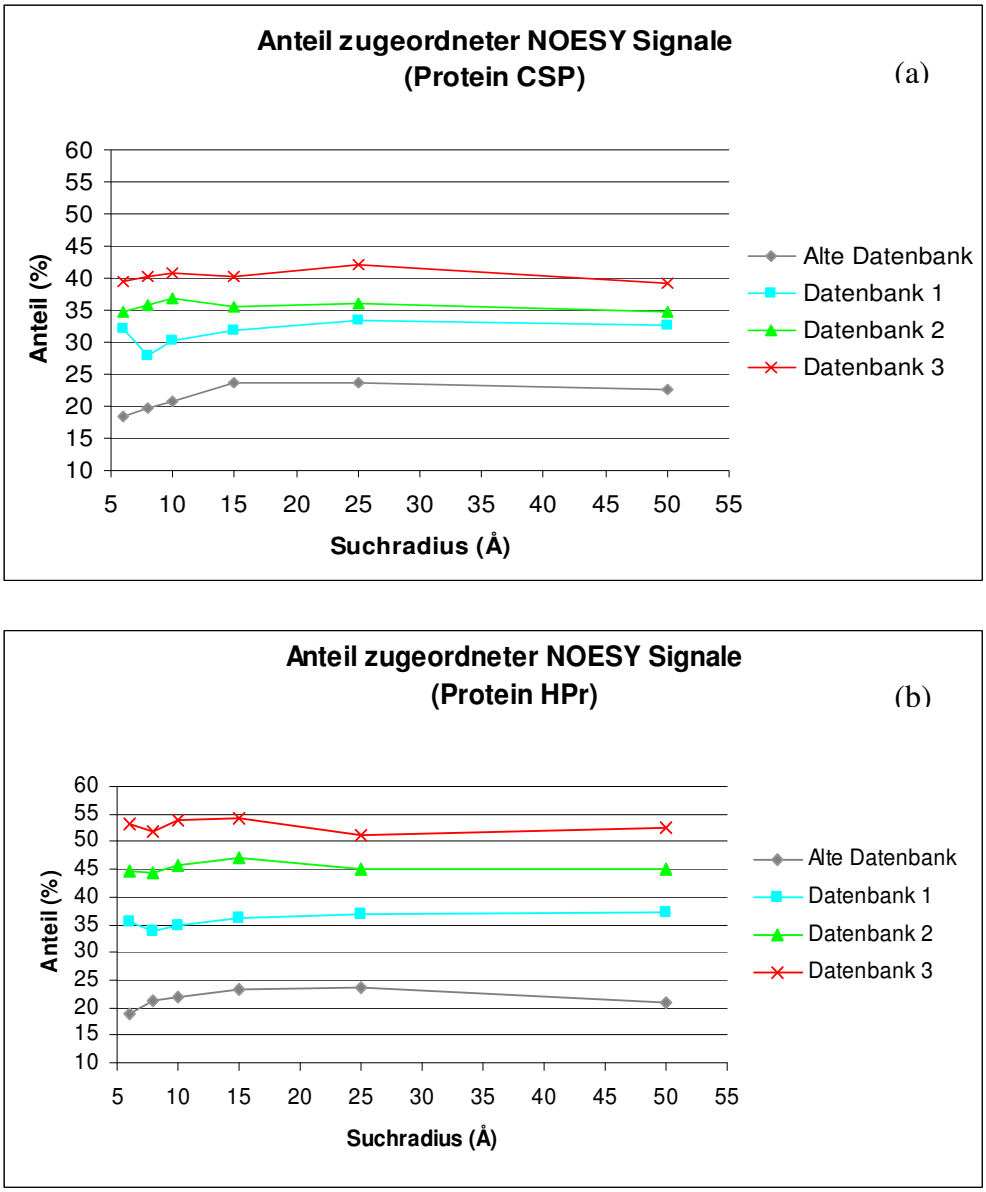


Abbildung 4.17: Anteile zugeordneter NOESY-Signale. Die Grafiken zeigen die prozentualen Anteile derjenigen zwei- und dreideutigen NOESY-Signale, für die eine der vorhandenen Zuordnungsmöglichkeiten eine Wahrscheinlichkeit von mindestens 98% berechnet werden konnte. Die Werte sind für unterschiedliche eingestellte Suchradien und unter Benutzung verschiedener Datenbanken für die NOESY-NMR-Spektren der Proteine CSP (oben) und HPr (unten) aufgeführt.

Diese Zuordnungsmöglichkeiten werden den NOESY-Signalen zugewiesen bzw. in die *Masterliste* übernommen. Das betreffende NOESY-Signal wird somit als eindeutig zugeordnet angesehen. In den gezeigten Grafiken sind die Werte für unterschiedliche benutzte Datenbanken bei verschiedenen eingestellten Suchradien dargestellt. Zur Orientierung ist in

den Grafiken noch zusätzlich die Gesamtzahl aller jeweils vorhandenen zwei- und dreideutigen NOESY-Signale dargestellt. Die Ergebnisse zeigen, dass man unter Benutzung der neuen Datenbanken von etwa ein Drittel (*Datenbank 1* und *2*) bis zu mehr als doppelt (*Datenbank 3*) so viele Signalzuordnungen, im Vergleich zur bisherigen Datenbank, erhielt. Das deutlich auffallende allgemeine Maximum der Zuordnungen beim Suchradius von 10 Å beruht, wie bereits in Kapitels 4.2.1.2 gezeigt, auf der entsprechend maximalen Anzahl vorhandener zwei- und dreideutige NOESY-Signale.
Ingesamt konnten die Anteile der zugeordneten zwei- und dreideutigen NOESY-Signale von etwa 20% auf etwa 40-55% gesteigert werden. Wie man sieht, sind die Werte weitgehend unabhängig vom eingestellten Suchradius.

4.2.2.2 Zunahme von Zuordnungen für unterschiedliche NOESY-Signale

Es hat sich gezeigt, dass die Anteile von zugeordneten *intraresiudualen, sequentiellen, mittelreichweitigen* und *langreichweitigen* zwei- und dreideutigen NOESY-Signalen sehr stark untereinander abweichen. In den Diagrammen der Abbildung 4.18 ist die Anzahl von *intraresidualen*, *sequentiellen*, *mittelreichweitigen* und *langreichweitigen* zugeordneten zwei- und dreideutigen NOESY-Signalen der simulierten 2D-NOESY-NMR Spektren von den Proteinen HPr und *Tm*CSP dargestellt. Die resultierenden Werte sind für die eingestellten Suchradien 0,6 nm und 10,0 nm aufgeführt. Die eingestellte Wahrscheinlichkeitsgrenze betrug P=0,98. Allgemein fällt auf, dass der Anteil von zugeordneten *intraresidualen* und *sequentiellen* Signalen, im Vergleich zu *mittelreichweitigen* und *langreichweitigen* NOESY-Signalen, stark überwiegt. Die zahlenmäßigen Verhältnisse der erstellten Zuordnungen zwischen diesen unterschiedlichen Gruppen von NOESY bleiben auch bei verschiedenen eingestellten Suchradien ungefähr gleich. Beim Vergleich der erhaltenen Werte unter Benutzung der bisherigen Datenbank mit der neuen *Datenbank 3* kann man sehen, dass insbesondere mehr *intraresiduale* und *sequentielle* Signale zugeordnet werden konnten. Bei *mittelreichweitigen* und *langreichweitigen* NOESY-Signalen hingegen lässt sich nicht in allen Fällen eine nennenswerte Steigerung von Zuordnungen beobachten.
In den Grafiken der Abbildung 4.19 sind die erreichten prozentualen Zuwächse an den unterschiedlichen NOESY-Signalgruppen unter Benutzung der *Datenbank 3* dargestellt. Es fällt auf, dass unter einem großen eingestellten Suchradius, im Vergleich zum kleinen Suchradius, wesentlich mehr zusätzliche Zuordnungen erzielt werden konnten. Dieses

Ergebnis war zu erwarten, da, wie bereits bekannt, mit kleiner werdenden Suchradius der Anteil zwei- und dreideutiger Signale stark abnimmt (s. Kap 4.2.1). Weiter fällt auf, dass für das NOESY-NMR-Spektrum vom Protein HPr im Vergleich zum Spektrum vom CSP generell ein größerer Zuwachs von Signalen erreicht werden konnte. Darauf wird in der Diskussion noch näher eingegangen.

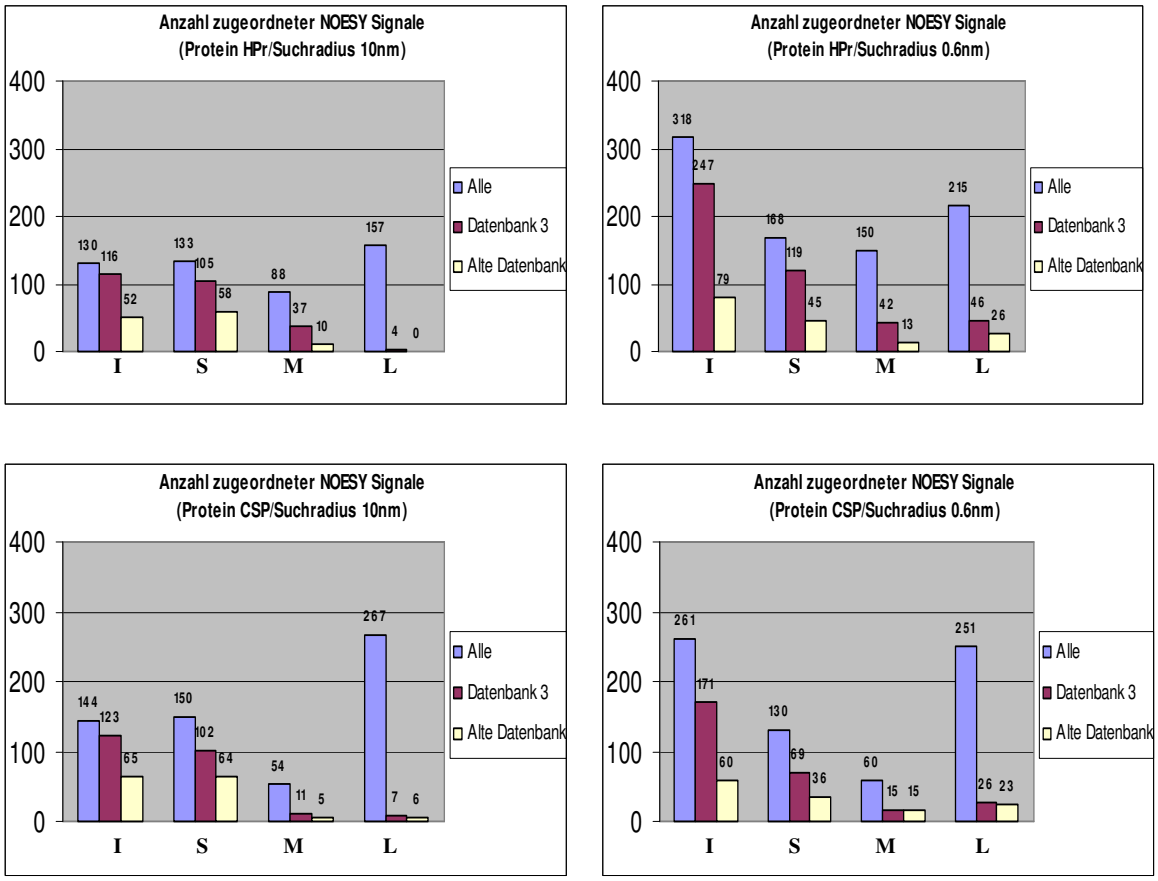


Abbildung 4.18: Anzahl erstellter Zuordnungen für unterschiedliche NOESY-Signale. In den Grafiken ist die Anzahl von zugeordneten *intraresiudualen* (I), *sequentiellen* (S), *mittel-reichweitigen* (M) und *langreichweitigen* (L) zwei- und dreideutigen NOESY-Signale aufgeführt. Jede Grafik zeigt die Werte für einen bestimmten eingestellten Suchradius (0,6 nm und 10,0 nm) unter Benutzung der früheren Wahrscheinlichkeitsverteilungen (weis) und der neuen *Datenbank 3* (violett). Die Ergebnisse sind für die simulierten 2D-NMR -Spektren der Proteine HPr (oben) und CSP(unten) aufgeführt. Zum Vergleich ist die Gesamtmenge (blau) der bei den genannten Suchradien vorhandenen unterschiedlichen Arten von zwei- und dreideutigen NOESY-Signalen,. aufgeführt. Die eingestellte Wahrscheinlichkeitsgrenze betrug P=0,98.

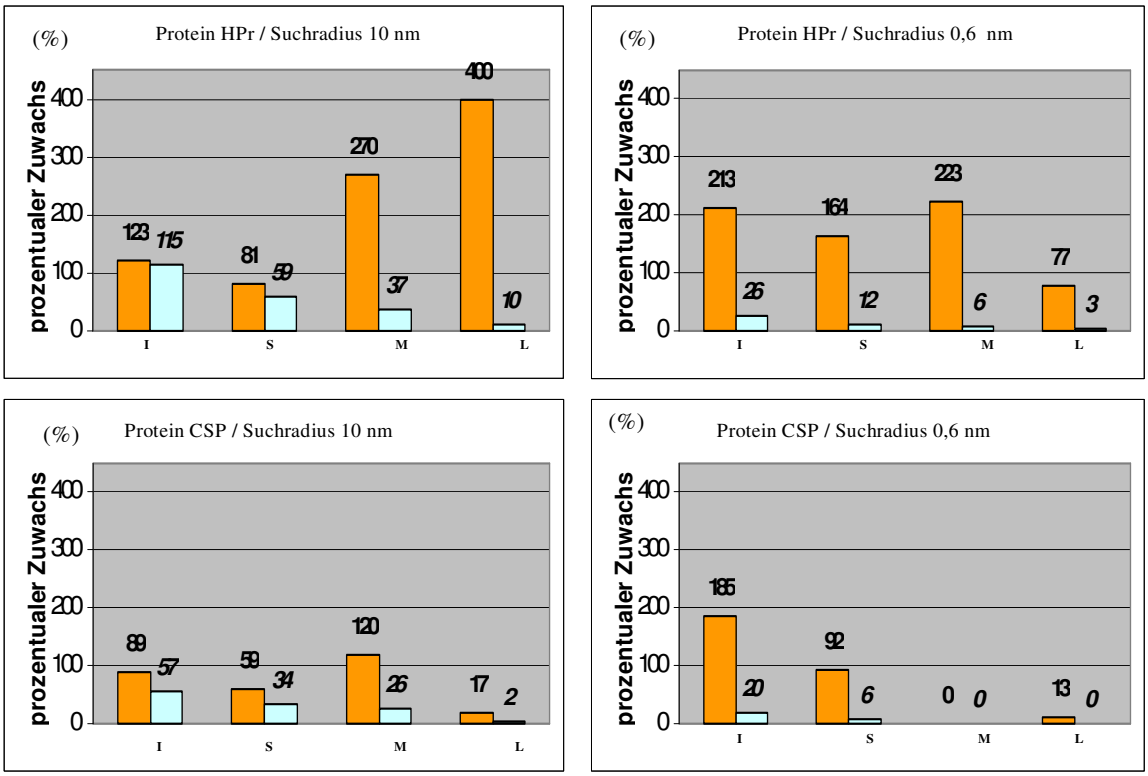
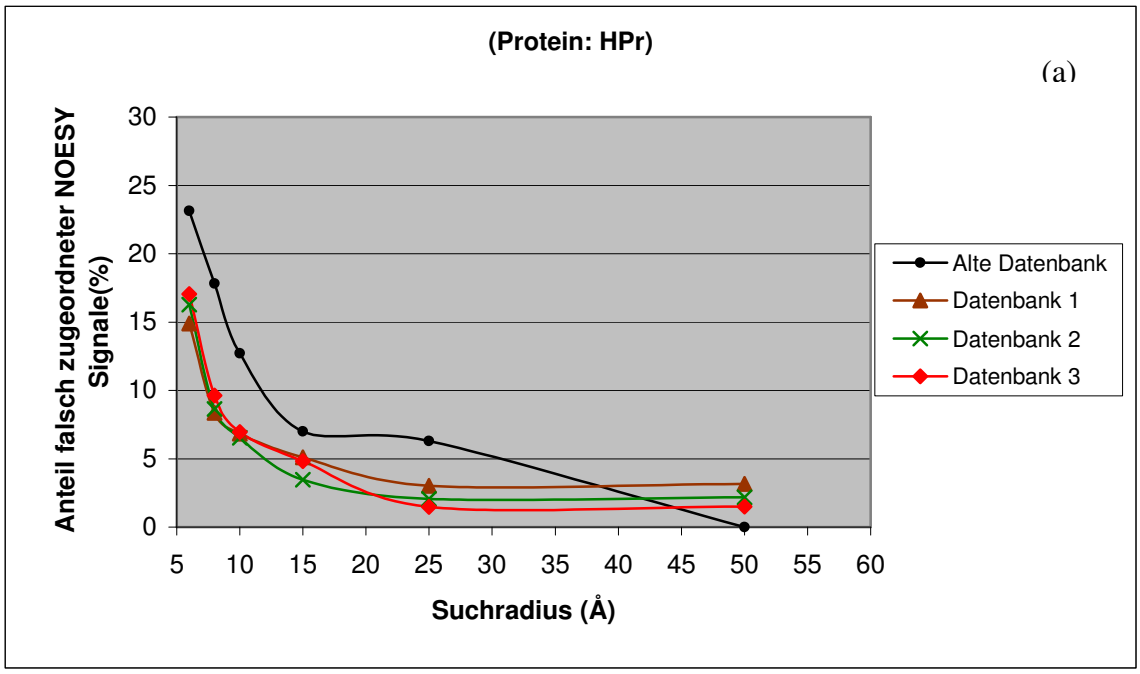


Abbildung 4.19: Prozentualer Zuwachs unterschiedlicher NOESY-Signale. In den Grafiken ist der prozentuale Zuwachs von unterschiedlichen zugeordneten NOESY-Signalen bei Benutzung der *Datenbank 3* aufgeführt. I steht für *intraresiduale,* S für *sequentielle*, M für *mittelreichweitige* und L für *langreichweitige* zwei- und dreideutige zugeordnete NOESY-Signale. Die orangefarbenen Balken stellen den erreichten Zuwachs von zwei- und dreideutigen zugeordneten Signalen gegenüber der Benutzung der bisherigen Datenbank dar. Die blauen Balken stellen den prozentualen Gesamtzuwachs der jeweiligen Signalgruppen da. Er bezieht sich auf alle, einschließlich eindeutig zugeordneten NOESY-Signalen der jeweiligen Signalgruppe unter Benutzung der bisherigen Datenbank. Die eingestellte Wahrscheinlichkeitsgrenze betrug P=0,98.

4.2.2.3 Reduktion falscher Zuordnungen

Hier soll gezeigt werden, in welchem Ausmaß unter Benutzung der neuen Wahrscheinlichkeitsdichteverteilungen der Anteil falscher Zuordnungen minimiert werden konnte. Die Diagramme in Abbildung 4.20 zeigen den prozentualen Anteil falsch zugeordneter zwei- und dreideutiger NOESY-Signale bei unterschiedlichen eingestellten Suchradien. Jede in den Diagrammen zu sehende Kurve zeigt die Werte für unterschiedliche benutzte Gruppen von Wahrscheinlichkeits(dichte)verteilungen. Das obere Diagramm in Abbildung 4.20 zeigt die Werte, die sich für das simulierte 2D-NOESY-NMR Spektrum des Proteins HPr ergeben haben, während das untere Diagramm die Werte für das 2D-NOESY-NMR-Spektrum vom Protein *Tm*CSP zeigt. Deutlich fällt auf, dass die Anteile falscher Zuordnungen mit abnehmendem Suchradius generell sehr stark zunehmen. Wie man allerdings sieht, fallen bei Benutzung der neuen *Datenbanken (1-3)* die Fehlerquoten



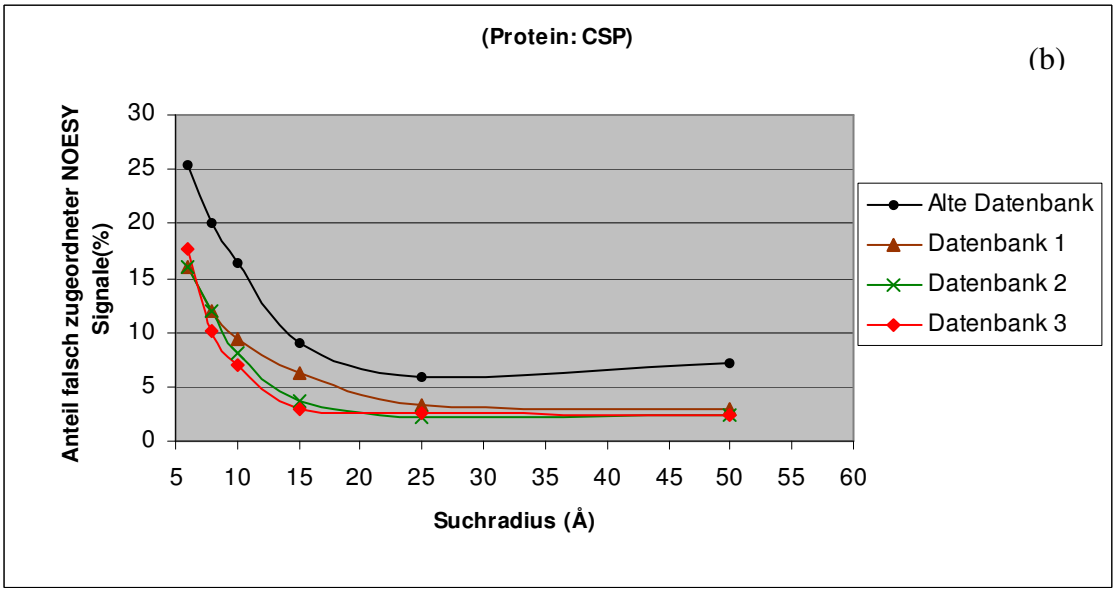


Abbildung 4.20: Falsch zugeordnete NOESY-Signale. Die Grafiken zeigen den prozentualen Anteil derjenigen zwei- und dreideutigen NOESY-Signale, welche mit einer mindestens 98% Wahrscheinlichkeit nicht die Zuordnung erhalten haben, die den größten Anteil am Signalvolumen bestimmt (=falsche Zuordnungen). Die Werte sind für die rückgerechneten 2D-NOESY-NMR Spektren der Proteine CSP (b) und HPr (a) für unterschiedliche eingestellte Suchradien aufgeführt. Jede Kurve zeigt die Werte bei Benutzung einer jeweils anderen Datenbank (s. Legende).Die eingestellte Wahrscheinlichkeitsgrenze betrug P=0,98.

im Vergleich zur Benutzung der bisherigen Wahrscheinlichkeitsverteilungen deutlich geringer aus. Die Fehlerquoten innerhalb den neuen Datenbanken (*Datenbanken 1-3)* sind in etwa vergleichbar.

4.2.2.4 Reduktion unerwünschter Zuordnungen

An dieser Stelle soll gezeigt werden, inwieweit, aufgrund der Benutzung der neuen Verteilungen, die Anteile von zugeordneten zwei- und dreideutigen NOESY-Signalen, deren Signalvolumen durch kein bestimmtes Atompaar zu mindestens 90 % erklärt wird (=unerwünschte Zuordnungen), minimiert werden konnte. Die Ergebnisse sind in Abbildung 4.21 für den relativ kleinen eingestellten Suchradius von 0,6 nm gezeigt, da hierbei die prozentualen Anteile dieser Signale unter den vorhandenen zwei- und dreideutigen Signalen besonders hoch sind.

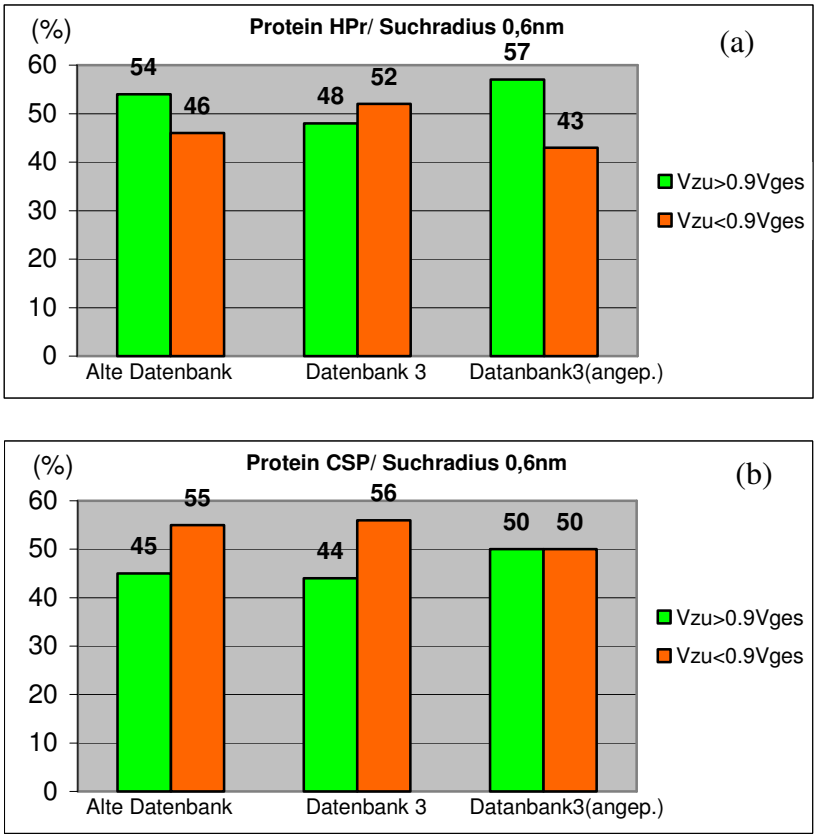


Abbildung 4.21: Reduktion unerwünschter Zuordnungen. Die grünen Balken zeigen den prozentualen Anteil an zugeordneten zwei und dreideutigen NOESY-Signalen (Vzu), die durch ein bestimmtes Atompaar zu mindestens 90 % erklärt werden. Die orangefarbenen Balken zeigen den entsprechenden anderen Anteil(bei dem das nicht der Fall ist). Die Werte sind für die rückgerechneten 2D-NOESY-NMR Spektren der Proteine CSP (b) und HPr (a) bei einem eingestellten Suchradius von 0,6nm dargestellt. Auf der linken Seite sind die Ergebnisse unter Benutzung der bisherigen Datenbank und in der Mitte bzw. Rechts für die neue *Datenbank 3* dargestellt. Die Werte auf der rechten Seite wurden bei zusätzlicher Berücksichtigung vom eingestellten Suchradius erzielt. Die eingestellte Wahrscheinlichkeitsgrenze betrug P=0.98.

Aus den Diagrammen der Abbildung 4.21 ist zu entnehmen, dass im Allgemeinen das Verhältnis von unerwünschten zu erwünschten Zuordnungen in etwa gleich ist.
Ein wesentlicher Unterschied bei der Anwendung der bisherigen Datenbank und der neuen *Datenbank 3* ist nicht zu erkennen. Allerdings konnte das Zahlenverhältnis zwischen erwünschten und unerwünschten Zuordnungen wie auch zwischen richtigen und falschen Zuordnungen (s. folgendes Kapitel) durch zusätzliche Berücksichtigung vom eingestellten Suchradius positiv beeinflusst werden. Dies wurde bisher noch nicht gemacht. Der Grundgedanke beruht hierbei auf zwei Annahmen:

1. Wenn ein Benutzter einen Suchradius von z.B. 0,6 nm einstellt, wird davon ausgegangen, dass alle Zuordnungsmöglichkeiten mit einen Abstand von größer als 0,6 nm innerhalb der Teststruktur nicht wesentlich zum fraglichen NOESY-Signal beitragen.

2. Zusätzlich wird davon ausgegangen, dass sich alle übrigen Zuordnungsmöglichkeiten von weniger als 0,6 nm in der Teststruktur auch in der wirklichen Struktur in diesem Abstands bereich befinden.

Diese Annahmen stellen eine zusätzlich Information da, welche man bei der Berechnung für die wahrscheinlichste Zuordnung einfließen lassen kann und somit zu einem besseren Ergebnis führen sollten. Die zusätzliche Berücksichtung vom Suchradius wurde hierbei durch Erzeugung einer Datenbank aus Wahrscheinlichkeitsdichteverteilungen bewerkstelligt, in denen nur Abstände von weniger als 0,6 nm integriert wurden. Zusätzlich wurden die Integrationsgrenzen der *a priori Faktoren* (s. Kapitel 2.5.3 /Formeln 2.15 und 2.16) auf den für 0,6 nm entsprechenden Volumenwert von etwa 2.14×10^{-5} Å^3 beschränkt. Dies ist nicht unbedingt nötig gewesen, da die Werte für Bereiche von größer als 0,6 nm hierbei sehr schnell gegen Null gehen.
Aus den Diagrammen der Abbildung 4.21 ist zu entnehmen, dass, durch den Einsatz der neuen *Datenbank 3* mit Berücksichtigung des Suchradius, das Verhältnis zwischen erwünschten und unerwünschten Zuordnungen, gegenüber den bisherigen Verteilungen, leicht verbessert werden konnte.

4.2.2.5 Zusammenhang zwischen unerwünschten und falschen Zuordnungen

Eine genaue Analyse der falsch zugeordneten zwei- und dreideutigen NOESY-Signale hat ergeben, dass sich vor allem falsche Zuordnungen bei solchen Signalen häufen, die durch kein bestimmtes Atompaar zu mindestens 90 % erklärt werden (s. Abb. 4.22). Auf das Phänomen wird in der Diskussion noch näher eingegangen.

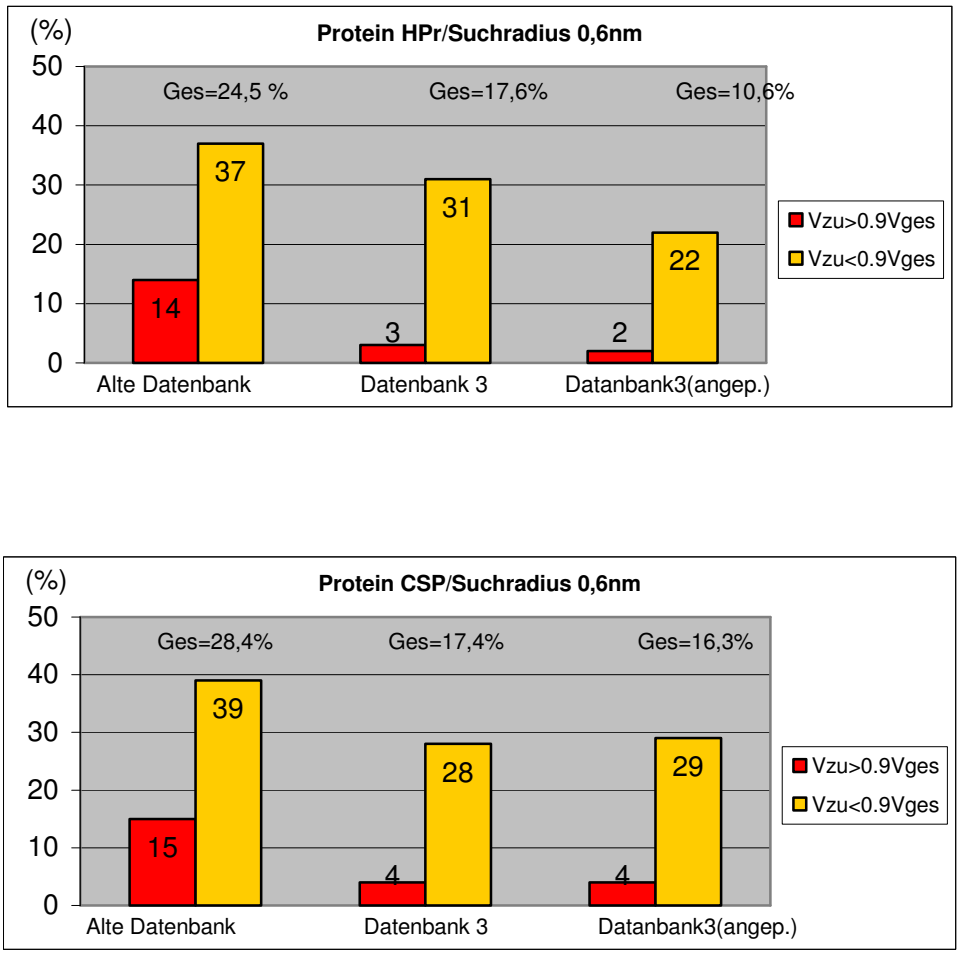


Abbildung 4.22: Anteile von Fehlzuordnungen für unterschiedliche NOESY-Signale. Die Diagramme zeigen den prozentualen Anteil falscher Zuordnungen für unterschiedliche Gruppen zugeordneter zwei- und dreideutiger NOESY-Signale (Vzu). Die roten Balken stehen jeweils für den Anteil falsch zugeordneter Signale, die durch ein bestimmtes Atompaar zu mindestens 90% erklärt werden. Die orangefarbenen Balken stehen für die entsprechend andere Gruppe. Über den Balken ist die Gesamtfehlerquote aller hierbei zugeordneter zwei- und dreideutiger NOESY-Signale aufgeführt. Das Gesamtvolumen eines Signals wird hier mit Vges bezeichnet. Die Ergebnisse sind bei Anwendung der bisherigen Verteilungen (links), der neuen *Datenbank 3 (*mitte) und der *Datenbank 3* unter Berücksichtigung vom Suchradius, dargestellt (rechts). Die eingestellte Wahrscheinlichkeitsgrenze betrug P=0,98.

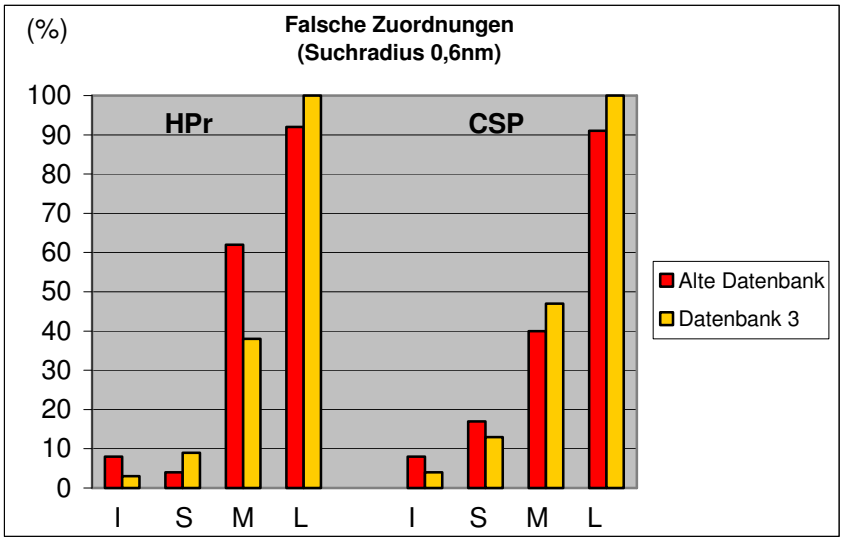
Wie die Diagramme in Abbildung 4.22 zeigen, konnte der Anteil an falschen Zuordnungen für NOESY-Signale, die durch ein bestimmtes Atompaar zumindest 90% erklärt werden, besonders stark minimiert werden. Aber auch für die andere Gruppe (=unerwünschte Zuordnungen) konnte eine deutliche Verringerung an Falschzuordnungen erreicht werden. Die Verringerung der Fehlerquoten für die erste Gruppe ist jedoch als wichtiger einzustufen, da hier Fehlzuordnungen im allgemeinen zu besonders großen Abstandsfehlern führen (s. Kap. 5.2.3.1).

4.2.2.6 Häufigkeit falscher Zuordnungen bei verschiedenen Arten von NOESY-Signalen

Es hat sich gezeigt, dass die Häufigkeit von falschen Zuordnungen unter *intraresidualen, sequentiellen*, *mittelreichweitigen* und *langreichweitigen* zwei- und dreideutig zugeordneten NOESY-Signalen stark unterschiedlich ist (s. Abb. 4.23). Die sich ergebenden Werte sind hier für relativ kurze Suchradien (0,6 nm und 1,0 nm) dargestellt, da hierbei potentiell die meisten Falschzuordnungen auftreten. Es ist zu erkennen, dass die Fehlerdichte bei *mittel-* und *langreichweitigen* NOESY-Signalen, im Vergleich zu *intraresidualen* und *sequentiellen* Signalen, auffallend stark erhöht ist. Insbesondere langreichweitige NOESY-Signale werden praktisch immer falsch zugeordnet.

Wenn man die Ergebnisse der früheren mit der neueren Datenbank vergleicht, stellt man fest, dass für *intraresiduale*, *sequentielle* und *langreichweitige* NOESY-Signale in etwa vergleichbare Werte erzielt wurden. Allerdings fällt auf, dass, unter Benutzung der *Datenbank 3* die Fehlerquote für *langreichweitige* NOESY-Signale immer bei 100% liegt. Für *mittelreichweitige* NOESY- Signale erreicht man unter Benutzung der *Datenbank 3* meist eine deutlich geringere Fehlerquote (Ausnahme: CSP bei Suchradius 0,6 nm).

Eine genauere Untersuchung der langreichweitigen zugeordneten zwei- und dreideutigen NOESY-Signale hat gezeigt, dass jeweils immer das sequentiell näher liegende Atompaar der vorhandenen Zuordnungsmöglichkeiten als Zuordnung gewählt wurde. Dies war zu erwarten, da Atompaare mit relativ kurzen sequentiellen Abständen grundsätzlich eine höhere statistische Wahrscheinlichkeit, gegenüber sequentiell weiter auseinanderliegenden Atompaaren, haben sich innerhalb eines Proteins räumlich nahe zu kommen (s. Kap. 4.1.2.3).



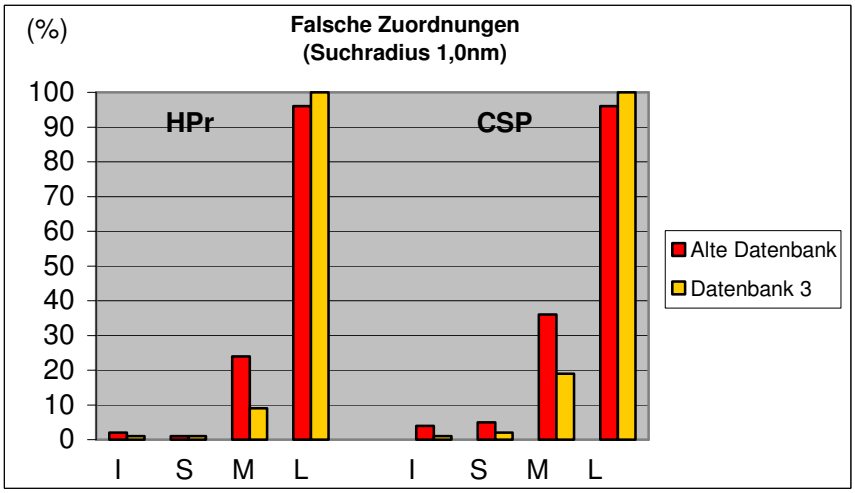


Abbildung 4.23: Verteilung falsch zugeordneter NOESY-Signale. Die Grafiken zeigen die prozentualen Anteile falscher Zuordnungen unter *intraresidualen* (I), *sequentiellen* (S), *mittelreichweitigen* (M) und *langreichweitigen* (L) zwei- und dreideutigen NOESY -Signalen. Die sich ergebenden Werte sind für die eingestellten Suchradien 0,6nm und 1,0nm, unter Benutzung der neuen *Datenbank 3* (orangefarbene Balken) und der herkömmlichen Datenbank (rote Balken), aufgeführt. Die Diagramme zeigen jeweils die Ergebnisse für die simulierten 2D-NOESY-NMR Spektren der Proteine *Tm*CSP (links) und HPr (rechts).

4.2.2.7 Verringerung des Abstandsfehlers

Wie bereits gezeigt wurde, konnte bei Anwendung der *neuen Datenbank 3* der Anteil der unerwünschten oder falsch zugeordneten zwei- und dreideutigen NOESY-Signale minimiert werden. Es somit zu erwarten, dass man entsprechend geringere Anteile an interatomaren Abständen mit entsprechend großen Abstandsfehlern bezüglich der wirklichen Struktur erhält.

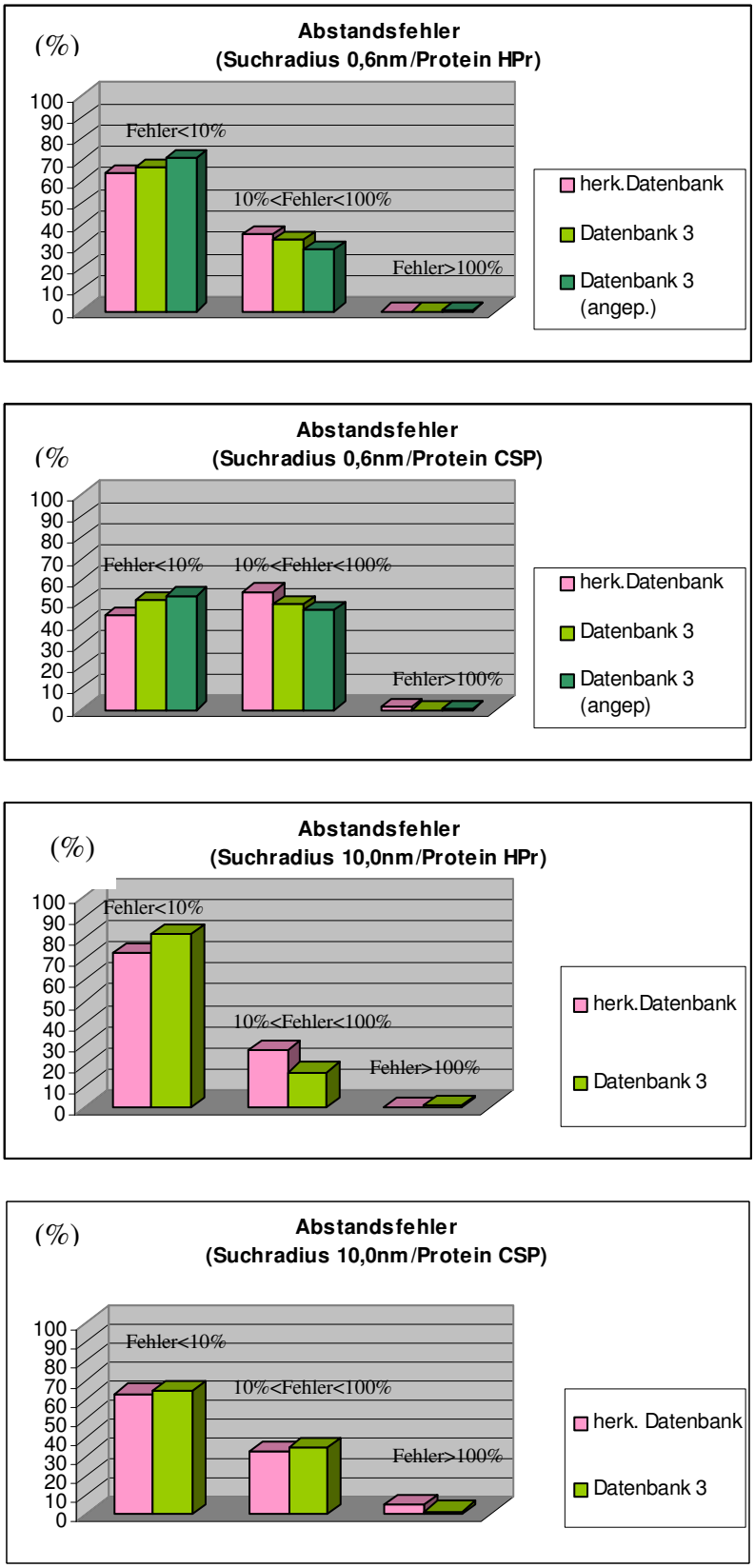


Abbildung 4.24: Minimierung vom Abstandsfehler. Die Diagramme zeigen den prozentualen Anteil von zwei- und dreideutigen zugeordneten NOESY-Signalen bei denen eine Abstands-abweichung von weniger als 10% (links),von 10% bis 100% (mitte) und mehr als 100% (Rechts) vom Abstand des zugeordneten Atompaares in der Struktur ermittelt wurde. Die Ergebnisse sind bei Benutzung der herkömmlichen Datenbank(rosa), der neuen *Datenbank 3* (hellgrün) und unter der neuen *Datenbank 3* bei jeweils gleichzeitiger Berücksichtigung vom Suchradius (dunkelgrün) (hier für 0,6nm) dargestellt. Die eingestellte Wahrscheinlichkeitsgrenze betrug P= 0.98.

Die Grafiken in der Abbildung 4.24 zeigen die prozentuale Verteilung von zugeordneten zwei- und dreideutigen NOESY-Signalen für jeweils unterschiedliche Größen des Fehlers bezüglich des ermittelten Abstandes. Der prozentuale Abstandsfehler wurde hierbei aus der Differenz des aus dem NOESY Volumen ermittelten Abstandes und dem wirklichen Abstand des zugeordneten Atompaares innerhalb der Struktur berechnet.
Generell fällt auf, dass bei einem kleinen Suchradius, im Vergleich zu einem großen Suchradius, der Anteil der ermittelten Abstände mit einer Abweichung von kleiner als 10% vom wirklichen Abstand kleiner ausfällt. Dies war zu erwarten, da, wie bereits gezeigt werden konnte, mit abnehmenden Suchradius der Anteil unerwünschter und falsch zugeordneter zwei- und dreideutiger NOESY-Signale deutlich ansteigt. Im Vergleich bei der Benutzung der neuen *Datenbank 3* gegenüber der früheren Datenbank zeigt sich meist eine deutliche Zunahme des Anteils von Abständen mit einer geringen Abweichung (<10%) bzw. Abnahme der Anteile von Abständen mit größeren Abstandsfehlern.
Die Ergebnisse zeigen weiter, dass der Anteil von ermittelten Abständen mit Abweichungen von mehr als 100% vom wirklichen Abstand generell eine sehr geringe Häufigkeit aufweisen (meist <2%). Bei Benutzung der früheren Datenbank fällt allerdings eine vergleichsweise starke Häufung von 5% dieser Abstände bei einem Suchradius von 10 nm für das simulierte Spektrum vom Protein CSP auf (Abb. 4.24 unten rechts). Hierbei wurden den betreffenden (sequentiellen) NOESY-Signalen fälschlicherweise Atompaare mit jeweils sehr großen sequentiellen Abständen von mehr als 20 Aminosäuren zugeordnet. Dies führte zu den großen Abstandsfehlern. Diese Art von falschen Zuordnungen sind besonders gefährlich, da sie während der Strukturrechnung die Tertiärstruktur des Proteins stark verzerren können. Unter Benutzung der neuen Datenbank wurden bei allen hier durchgeführten Testreihen solche Zuordnungen nicht beobachtet. Dies hat zwei Gründe:

1. Generell haben sequentiell nahe liegende Atompaare eine statistisch höhere Wahrscheinlichkeit innerhalb einer Proteinstruktur sich räumlich nahe zu kommen als vergleichsweise sequentiell weiter entfernt liegende Atome. Aufgrund dieser Tatsache, wird grundsätzlich dem in Frage stehenden NOESY-Signal das sequentiell näher liegende Atompaar der jeweils vorhandenen Zuordnungsmöglichkeiten als Zuordnung zugewiesen.

2. Die Wahrscheinlichkeitsdichteverteilungen von sequentiell relativ weit entfernter (etwa mehr als 5 Aminosäuren) unterschiedliche Atompaare weisen nur geringe Unterschiede in ihrem Kurvenverlauf auf (s. Kap 4.1.2.3).

Deshalb kann im Fall von mehreren Zuordnungsmöglichkeiten mit jeweils sequentiell relativ weit auseinanderliegenden Atompaaren keine Entscheidung gefällt werden. Aus diesen beiden Gründen konnte unter Benutzung der neuen *Datenbank 3* praktisch keine langreichweitige richtige oder falsche Zuordnung bei einem zwei- oder dreideutigen NOESY- Signal beobachtet werden.

4.2.2.8 Die Bedeutung der Wahrscheinlichkeitsgrenze

Alle bisher gezeigten Ergebnisse wurden bei der konstant eingestellten Wahrscheinlichkeitsgrenze von P=0,98 erzielt. Hier soll untersucht werden, wie sich unterschiedliche Werte der Wahrscheinlichkeitsgrenze auf die Zuordnungsqualität auswirken.
Generell ist bei einer Abnahme der Wahrscheinlichkeitsgrenze mit einer Zunahme von Zuordnungen zu rechnen, da nun auch Zuordnungen mit entsprechend geringeren ermittelten Wahrscheinlichkeitswerten zugelassen werden. Wie in Abbildung 4.25 zu sehen, steigt die Anzahl der zugeordneten zwei- und dreideutigen NOESY- Signale mit abnehmender Wahrscheinlichkeitsgrenze in etwa linear an. Allgemein erhöht sich die Anzahl, unabhängig von der jeweils benutzten Datenbank oder dem jeweiligen Spektrum, angefangen für P=0,99 bis P=0,6, in etwa um zwei Drittel. Die Ergebnisse sind hier für den eingestellten Suchradius von 1,0 nm aufgeführt. Mit Ausnahme der größeren Werte, zeigen die Kurven für die neue *Datenbank 3*, im Vergleich zur herkömmlichen Datenbank, keine wesentlichen Unterschiede im Kurvenverlauf oder des Steigungsverhaltens.
Abbildung 4.26 zeigt den prozentualen Anteil falsch zugeordneter zwei- und dreideutiger NOESY-Signale in Abhängigkeit von der jeweils eingestellten Wahrscheinlichkeitsgrenze. Allgemein kann man sehen, dass mit zunehmenden Werten der Anteil falscher Zuordnungen deutlich absinkt. Dies war aufgrund der Theorie zu erwarten und zeigt, dass das hier angewandte statistische Verfahren zur Zuordnung von zwei- und dreideutigen NOESY-Signalen prinzipiell funktioniert. Wie man weiter sieht, fallen die Fehlerquoten für das 2D-NOESY-Spektrum vom HPr, gegenüber vom Protein CSP, im Allgemeinen geringer aus. Sie nähern sich aber mit steigender Wahrscheinlichkeitsgrenze, unter Benutzung der jeweils gleichen Datenbank, immer stärker aneinander an.
Beim Vergleich der Kurven fällt weiter auf, dass unter Benutzung der neuen *Datenbank 3,* neben den generell geringern Fehlerquoten, die Werte mit zunehmender Wahrscheinlichkeitsgrenze stärker abnehmen. Entsprechend divergieren die Kurven für die neue bzw. herkömmliche Datenbank immer weiter auseinander.

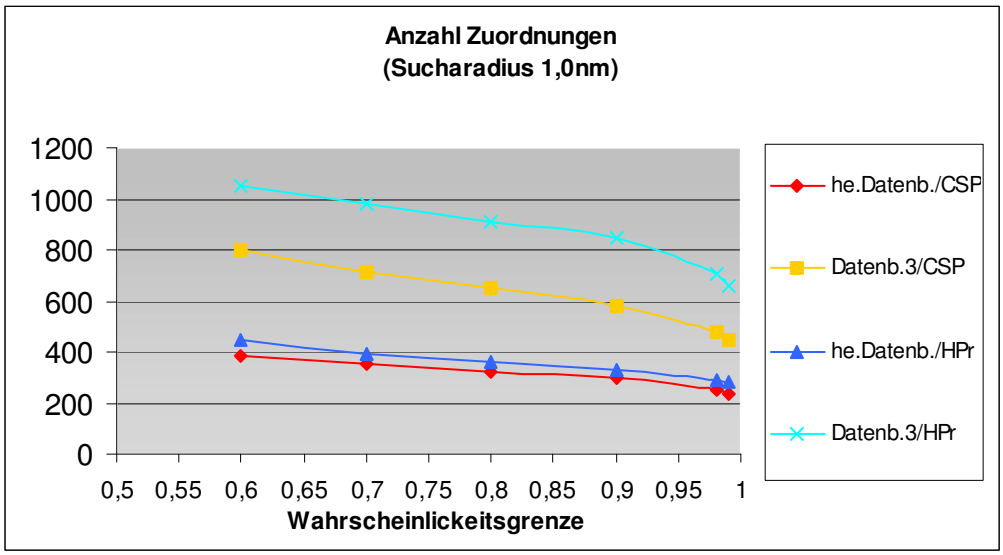


Abbildung 4.25: Rolle der Wahrscheinlichkeitsgrenze auf die Zuordnungsanzahl. Die Grafik zeigt die Abhängigkeit der Anzahl zugeordneter zwei- und dreideutiger Zuordnungen von der eingestellten Wahrscheinlichkeitsgrenze (für P=0.99,0.98,0.9.. 0.6) Die Werte sind für die simulierten 2D-NOESY-NMR-Spektren der Proteine HPr und CSP unter Benutzung der herkömmlichen Datenbank und der neuen *Datenbank 3* dargestellt.

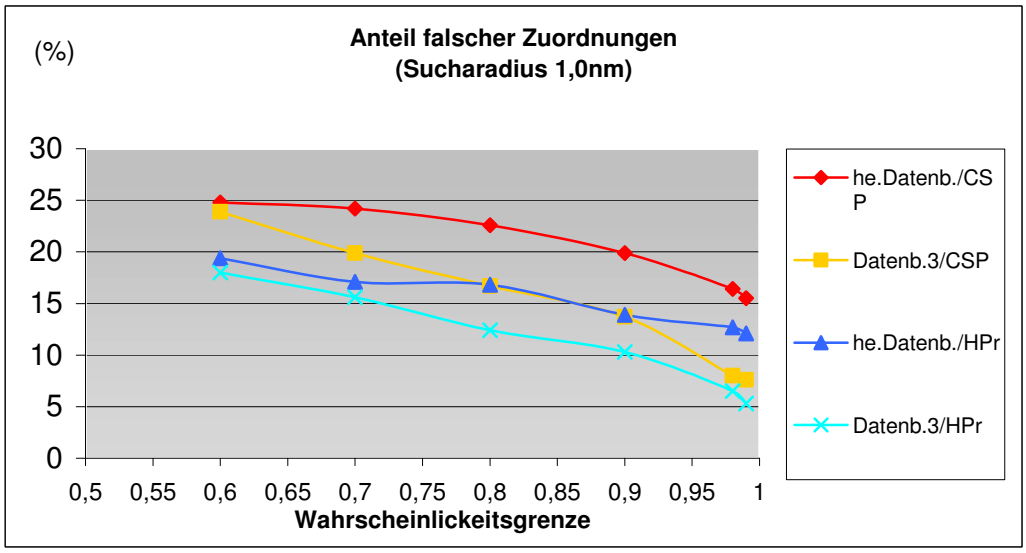


Abbildung 4.26: Rolle der Wahrscheinlichkeitsgrenze auf die Anteile falscher Zuordnungen. Die Grafik zeigt die Abhängigkeit der Anteile falsch zugeordneter zwei- und dreideutiger Zuordnungen von der eingestellten Wahrscheinlichkeitsgrenze (für P=0.99, 0.98, 0.9.. 0.6). Die Werte sind jeweils für die simulierten 2D-NOESY-NMR-Spektren der Proteine HPr und *Tm*CSP unter Benutzung der herkömmlichen Datenbank und der neuen *Datenbank 3* dargestellt.

Allgemein lässt sich sagen, dass mit steigender Wahrscheinlichkeitsgrenze, unter Abnahme der Gesamtzuordnungen, die Zuordnungssicherheit zunimmt. Dies sieht man an der Verkleinerung der Anteile falscher Zuordnungen sowie der Minimierung der Unterschiede der Fehlerquote zwischen den unterschiedlichen Proteinen. Allerdings fallen die Ergebnisse, unter

Benutzung der neuen *Datenbank 3* im Vergleich zur herkömmlichen Datenbank, mit steigender Wahrscheinlichkeitsgrenze zunehmend besser aus.

4.2.2.9 Bedeutung der Datenauflösung

Hier soll untersucht werden, inwieweit die Datenauflösung der benutzten Wahrscheinlichkeitsdichteverteilungen (hier *Datenbank 3*) sich auf die Anzahl der zugeordneten zwei- und dreideutigen NOESY-Signale auswirkt. Hierfür wurden, unter Benutzung von Volumenwahrscheinlichkeitsdichteverteilungen unterschiedlicher Datenauflösungen (je 10000, 1000, 500, 250, 125 und 64 Datenpunkten), mit dem Programm *KNOWNOE* Zuordnungen der simulierten 2D-NOESY-NMR-Spektren von den Proteinen CSP und HPr erstellt. Aus der Grafik in Abbildung 4.27 ist zu entnehmen, dass mit sinkender Datenauflösung die Anzahl der mit jeweils einer hohen Wahrscheinlichkeit (98%) zugeordneten zwei- und dreideutigen NOESY-Signale deutlich abnimmt.

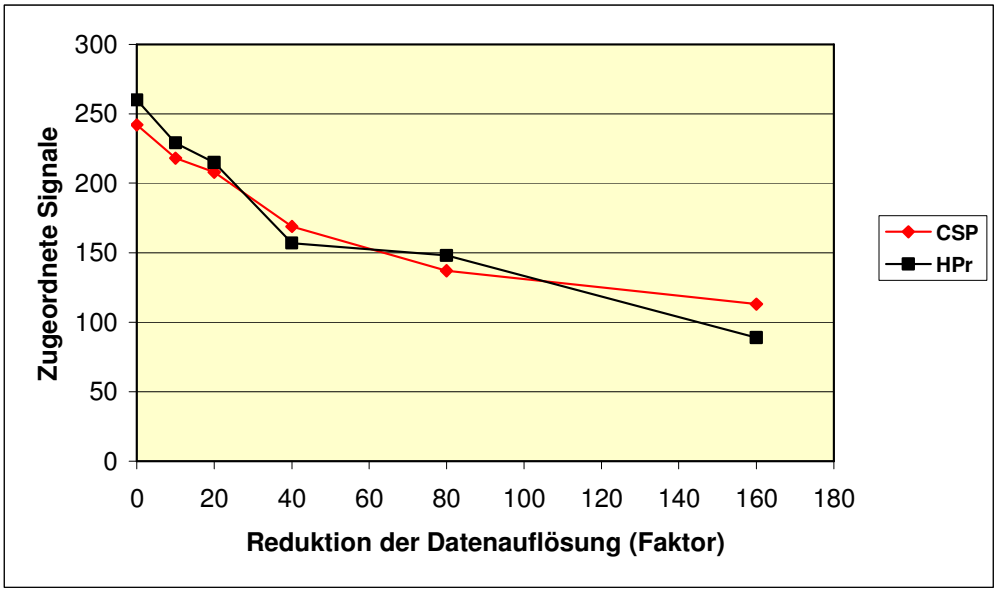


Abbildung 4.27: Die Bedeutung der Datenauflösung. Die Grafik zeigt die Anzahl zugeordneter zweideutiger zugeordneter NOESY-Signale. Der Suchradius betrug 10 nm. Die Werte sind für die rückgerechneten 2D-NOESY-NMR Spektren der Proteine CSP (rot) und HPr (schwarz) aufgeführt. Die eingestellte Wahrscheinlichkeitsgrenze war P=0.98.

Wie man weiter sieht, ist die Abnahme der Zuordnungen weitgehend unabhängig vom Spektrum bzw. dem zugehörigen Protein. Sie verläuft in etwa mit der Zunahme des Faktors der Reduktion der Datenauflösung linear. Dies zeigt, dass die hier ausgewählte hohe Auflösung der Verteilungen nicht redundant war bzw. zu einem Gewinn an struktureller Abstandsinformation führte.

4.2.2.10 Einfluss falscher Abstände

In der Praxis ist die Abstandsbestimmung aus NOESY-Signalen stets mit Fehlern behaftet. Diese Tatsache hat allerdings auch Auswirkungen auf die Zuordnungsqualität der Spektren. Ein falsch ermittelter Abstand führt hierbei zur Berechnung entsprechend falscher Integralbereiche innerhalb der Volumenwahrscheinlichkeitsdichteverteilungen. Dadurch werden wiederum die Wahrscheinlichkeiten der vorhandenen Zuordnungsmöglichen für ein entsprechend abweichendes Signalvolumen berechnet. Dies kann die Zuordnungsqualität in zweierlei Hinsicht negativ beeinflussen:

Zu einem kann es passieren, dass sich, aufgrund des falschen Abstandes, eine höhere Wahrscheinlichkeit für die falsche Zuordnungsmöglichkeit ergibt. Somit besteht die Gefahr einer erhöhten Fehlerquote. Anderseits kann auch genauso gut der umgekehrte Fall eintreten, bei dem für die richtige Zuordnungsmöglichkeit eine höhere Wahrscheinlichkeit berechnet wird. Es ist anzunehmen, dass sich beide Effekte gegenseitig ausgleichen werden und somit insgesamt zu keiner nennenswerten Veränderung der Fehlerquote kommen dürfte. Es ist eher zu erwarten, dass, aufgrund Fehler bei der Abstandsberechnung, die Wahrscheinlichkeiten (in Wirklichkeit) für alle der jeweils vorhandenen Zuordnungsmöglichkeiten der NOESY-Signale in der Regel geringer werden und damit insgesamt weniger Signale zugeordnet werden können. Zur Überprüfung des Einflusses falscher Abstände auf die Zuordnungsqualität wurden jeweils Testreihen mit künstlich eingebauten Abstandsfehlern erstellt (s. Abb. 4.28). Hierbei wurden die aus den NOESY-Signalen ermittelten Abstände, durch Abstände ersetzt, welche einen definierten Prozentsatz (je + / − 30%) von dem wirklichen Abstand der jeweils richtigen Zuordnung innerhalb der Proteinstruktur abweichen. Auf Basis dieser verfälschten Abstände, wurden dann die entsprechenden Wahrscheinlichkeiten der in Frage stehenden NOESY-Signale berechnet. Die hier durchgeführten Testreihen wurden für einen etwa mittelgroßen Suchradius (1,0 nm) durchgeführt. Wie man aus den Grafiken der Abbildung 4.28 entnehmen kann, erhält man unter Benutzung der neuen *Datenbank 3* in etwa die zu erwarteten Ergebnisse. So bleibt der Anteil falscher Zuordnungen, im Gegensatz zur Benutzung der herkömmlichen Datenbank, auch bei durchgängig falschen benutzten Abständen relativ konstant. Auch kann man generell eine Abnahme von Zuordnungen bei falschen Abständen verzeichnen. Nur beim 2D-NOESY-NMR-Spektrum vom Protein CSP nahm die Anzahl der Zuordnungen für die künstlich verlängerten Abstände etwa zu. Der Grund hierfür liegt vermutlich in der relativ kurzen Sequenzlänge von 66 Aminosäuren.

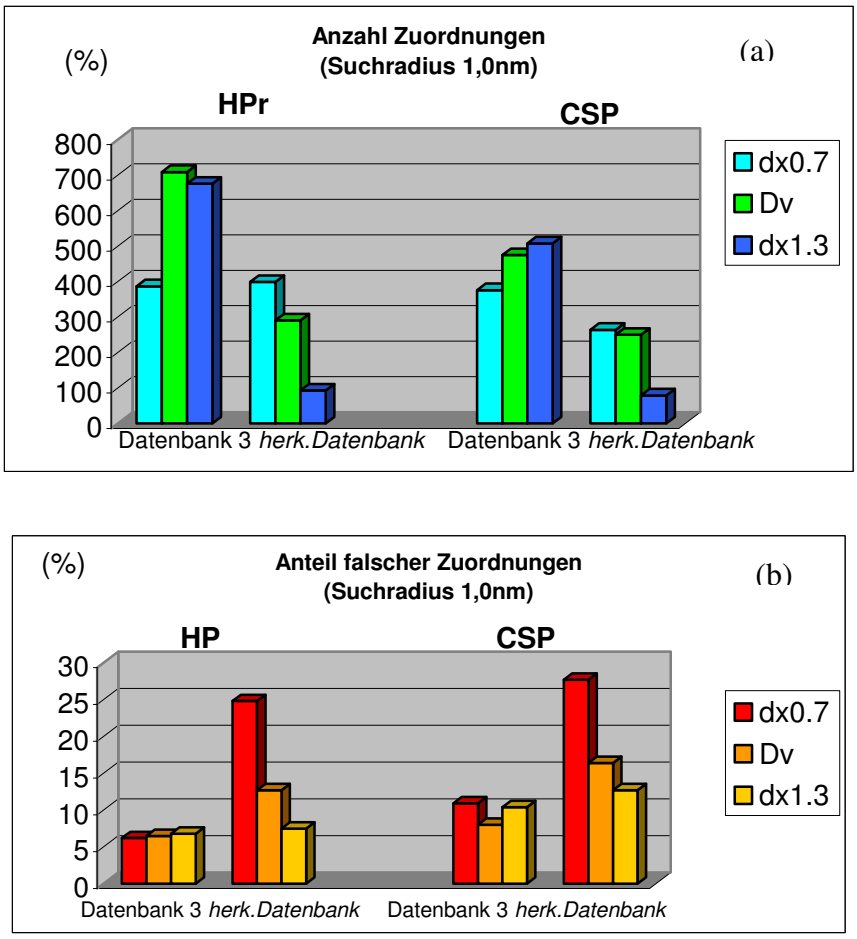


Abbildung 4.28: Einfluss von Abstandfehlern. Die Grafiken zeigen die Auswirkung von künstlich eingebauten Abstandsfehlern auf die Gesamtzahl (Grafik a) von zugeordneten zwei- und dreideutigen NOESY-Signalen so wie auf den prozentualen Anteil an falschen Zuordnungen (Grafik b). Die mittleren Balken (dunkelorange bzw. grün) einer Dreierreihe zeigen jeweils die Werte, die man für den aus dem Signalvolumen erhaltenen Abstand (Dv) erhalten hat (durchschnittlich etwa 10% negative Abweichung vom wirklichen Abstand der richtigen Zuordnungsmöglichkeit). Die hellblauen bzw. roten Balken zeigen die Werte für Abstände die um 30% kleiner sind als der Abstand der richtigen Zuordnung (=d) in der Proteinstruktur. Die dunkelblauen bzw. dunkelgelben Balken zeigen die Werte für um 30 % größere Abstände. Die eingestellte Wahrscheinlichkeitsgrenze betrug P=0.98. Eine Dreierreihe von Balken steht für die erzielten Werte unter Benutzung einer bestimmten Datenbank.

Da die durchschnittliche Sequenzlänge der in der *Datenbank 3* verarbeiteten Proteine mehr als viermal so groß ist (271 Reste), kommen im Vergleich zum CSP, anteilsmäßig für jedes Protein in der hier benutzen Strukturdatenbank mehr größere Abstände vor. Dies hat zur Folge, dass entsprechend größere Abstände im Durchschnitt als wahrscheinlicher angesehen werden, was wiederum zur erhöhten Anzahl von Zuordnungen führt.

Unter Benutzung der herkömmlichen Datenbank kommt es bei Verkleinerung der Abstände ‚unerwarteter Weise, bei jeweils beiden Spektren zu einer Zunahme von Zuordnungen. Die künstliche Abstandsvergrößerung hingegen, bewirkt eine unverhältnismäßig starke Abnahme an Zuordnungen. Allgemein kann man sagen, dass die hier angewandte wissensbasierte

Methode zur Zuordnung von zwei- und dreideutigen NOESY-Signalen unter Benutzung der neuen *Datenbank 3*, sich gegenüber falschen Abständen, wesentlich stabiler verhält.

4.2.2.11 Die Bedeutung des relativen Sequenzabstands bei der Bildung von Abstandsklassen

Wie bereits in Kapitel 4.1.1.2 erwähnt, wurden im Vergleich zu den bisherigen Verteilungen, noch zusätzliche Abstandsklassen für Atome mit jeweils 5,6,7, und 8 Aminosäuren Sequenzabstand gebildet. Grundlage hierfür war die Beobachtung, einer starken linearen Zunahme der mittleren räumlichen Abstände für Atome bis zu etwa 12 Aminosäuren Sequenzabstand (Abb. 4.8/ Kapitel 4.1.2.3). Deshalb wurde zunächst angenommen, dass sich die Verteilungen der neu gebildeten Abstandsklassen entsprechend stark voneinander unterscheiden werden und damit eine Zunahme von zugeordneten zwei- und dreideutigen NOESY-Signalen erwartet. Hier soll untersucht werden inwieweit sich eine Bildung von Abstandsklassen für Atome mit jeweils bestimmten sequentiellen Abständen auf die Zunahme von Signalzuordnungen auswirkt. Dabei wurde so vorgegangen, dass die Abstandklassen bezüglich des Sequenzabstandes, innerhalb der Datenbank 3, schrittweise um jeweils einen sequentiellen Abstand von einer Aminosäure entfernt wurden. Das heißt, dass es z.B. nach dem ersten Reduktionsschritt separate Verteilungen für Atome mit Sequenzabständen anstatt von 0-8 Aminosäuren Abstand, nur noch Verteilungen für jeweils von 0-7 Aminosäuren Abstand gibt. Abstände von Atomen mit sequentiellen Abständen von mehr als 7 Aminosäuren wurden dabei in jeweils eine Verteilung zusammengefasst. Ansonsten blieb die Art und Weise der Klassenbildung von Abständen unverändert. Im nächsten Reduktionsschritt wurde entsprechend analog vorgegangen. Die Datenbank wurde dabei soweit reduziert, bis es nur noch Verteilungen bzw. Abstandsklassen für intraresiduale Atome gab. Nach jedem Reduktionsschritt wurde die resultierende Datenbank bezüglich ihrer Leistungsfähigeit zwei- und dreideutige Signale zuzuordnen überprüft. Das Ergebnis ist inden Grafiken der Abbildung 4.29 zu sehen. Wie man sehen kann, spielt eine separate Abstandsklassenbildung für relative Sequenzabstände von 1-8 Aminosäuren bezüglich der Anzahl zugeordneter langreichweitiger zwei- und dreideutiger NOESY-Signale so gut wie keine Rolle. Für mittelreichweitige Signale konnte man für das Spektrum vom Protein HPr eine Steigerung der Zuordnungen mit einer separaten Abstandklassenbildung bis zu jeweils 4 Aminosäuren Abstand erreichen. Für sequentielle NOESY-Signale war insbesondere eine gesonderte Klassenbildung für Atome mit jeweils einer Aminosäure sequentiellen Abstand

für die Zuordnungsanzahl besonders wichtig. Auf die Anzahl zugeordneter intraresidualer NOESY-Signale wirkte sich eine separate Bildung von Abstandklassen für relative Sequenzabstände von größer als einer Aminosäure so gut wie gar nicht aus. Zusammenfassend haben die Ergebnisse gezeigt, dass sich für die Anwendung eine separate Bildung von Abstandklassen für Atome mit relativen sequentiellen Abständen bis zu maximal 4 Aminosäuren lohnt.

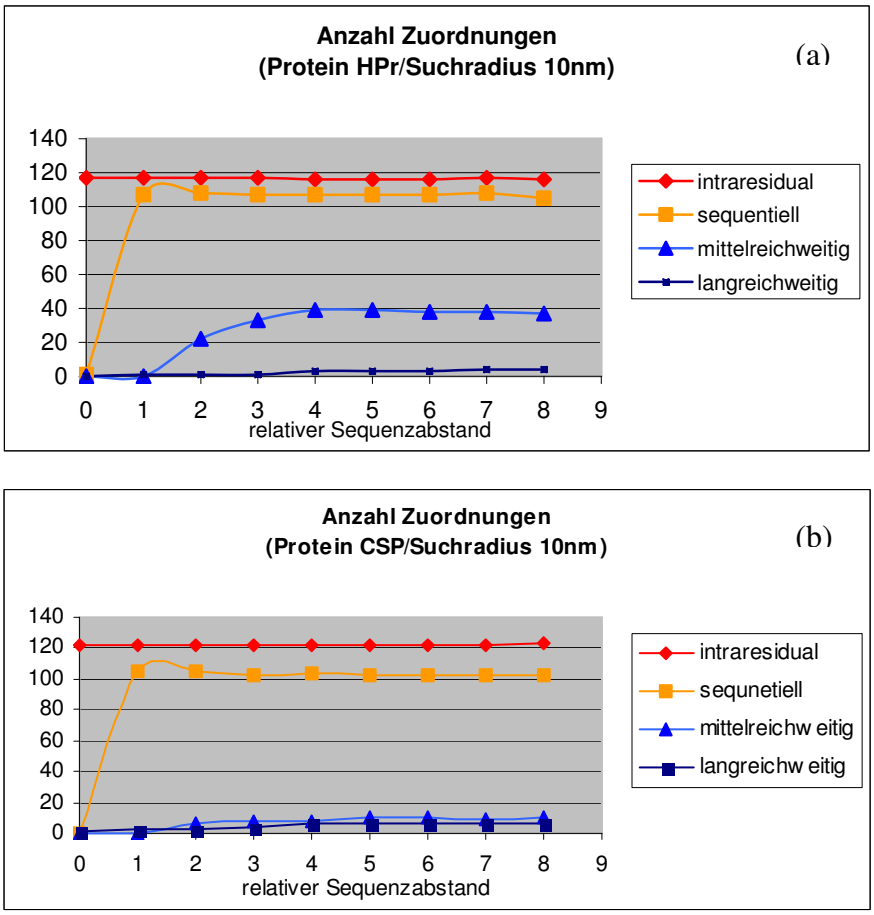


Abbildung 4.29 Rolle des sequentiellen Abstandes bei der Bildung von Abstandsklassen. Die Grafiken zeigen die erreichte Anzahl von zugeordneten intraresidualen, sequentiellen, mittelreichweitgigen und langreichweitigen zwei- und dreideutigen NOESY-Signalen. Die Werte sind jeweils unter Verwendung unterschiedlich stark erweiterter Datenbanken aufgeführt. So stehen beispielsweise am Punkt 3 der x-Achse die Werte unter Benutzung einer Datenbank, für die jeweils separate Verteilungen für Atome mit sequentiellen Abständen von 0,1,2 und 3 Aminosäuren erzeugt wurden. Die Versuche wurden hierbei bei einem Suchradius von 10nm und einer eingestellten Wahrscheinlichkeitsgrenze von P=0.98 durchgeführt. Die Grafik a zeigt die Ergebnisse für das 2D-NOESY-NMR Spektrum vom Protein HPr bzw. die Grafik b für das Protein CSP.

Kernpunkt der Arbeit war die Extraktion und Anwendung statistischer Information von atomspezifischen Abständen aus einer Vielzahl strukturell bekannter Proteine. Diese wurden in Form umfangreicher Datenbanken aus Abstands- oder Volumenwahrscheinlichkeitsdichte-verteilungen zu Verfügung gestellt. Die in den Datenbanken enthaltenden Informationen können zur Optimierung unterschiedlicher für die Proteinstrukturbestimmung benötigter

Arbeitschritte eingesetzt werden. Es konnte gezeigt werden, dass bei Verwendung der hier erzeugten Wahrscheinlichkeitsverteilungen die Leistungsfähigkeit des Programms *KNOWNOE,* dessen Aufgabe die automatische Zuordnung von NOESY-Spektren ist, deutlich verbessert werden konnte. Umgekehrt konnte anhand der Anwendung die Datenbanken optimiert, und aufgetretene Fehler bei ihrer Generierung aufgespürt werden. Aufgrund der Ergebnisse, konnte im Weiteren eine Einschätzung über den statistisch relevanten Informationsgehalt über interatomare Atomabstände innerhalb vieler strukturell bekannter Proteine geliefert werden. Diese kann zur Orientierung für den Einsatz der Datenbanken in zukünftigen Anwendungen herangezogen werden. Im ersten Teil der Diskussion soll auf die Verbesserung des Programms *KNOWNOE* eingegangen werden. Dabei wird insbesondere auf die Bedeutung der Ergebnisse für die Strukturbestimmung eingegangen werden. Im zweiten Teil der Diskussion soll anhand der Testergebnisse diskutiert werden, welche Faktoren bei der Generierung der Verteilungen besonders wichtig bzw. weniger wichtig gewesen sind. Des weiteren soll dabei erläutert werden, inwieweit strukturell bekannte Proteine statistisch relevante Abstandsinformationen enthalten.

5.1 Versuchsbedingungen

5.1.1 Testspektren

Die Überprüfung der Qualität der automatischen Zuordnung von NOESY-Signalen unter Einsatz der neuen Wahrscheinlichkeitsdichteverteilungen wurde mit den simulierten 2D-NOESY-NMR-Spektren der Proteine HPr und *Tm*CSP durchgeführt. Die Durchführung der Testreihen an zwei Spektren unterschiedlicher Proteine sollte zeigen, ob oder inwieweit die Zuordnungsqualität von den Eigenschaften des jeweils zu untersuchenden Proteins abhängig ist. Dies wiederum ließ Rückschlüsse auf den Grad der Unabhängigkeit der hier erzeugten Datenbanken bezüglich der Eigenschaften der zu untersuchenden Proteine zu.

Die Durchführung der Versuche an simulierten NOESY-NMR-Spektren war notwendig, da hierbei die richtige Zuordnung aller Signale bekannt ist, und somit erst eine Überprüfung der Richtigkeit der Zuordnung von den jeweils automatisch zugeordneten NOESY-Signalen möglich war.

In experimentellen NOESY-Spektren kommt es, im Gegensatz zu simulierten Spektren, aufgrund geringer gegenseitiger Unterschiede bezüglich der chemischen Verschiebung oft zur Verschmelzung von mehreren Signalen. Um den experimentellen Bedingungen näher zu kommen wurden deshalb simulierte NOESY-Signale, welche sich in beiden

Frequenzdomänen um weniger als 0,015 ppm unterscheiden, durch Aufsummierung zu einem Signal zusammengefasst.

Für die Überprüfung der maximalen Kapazität mit Hilfe der im Rahmen der Arbeit erzeugten Datenbanken Signale zuzuordnen war es wichtig, dass der aus den NOESY-Signalen ermittelte Abstand nicht zu stark von den wirklichen Abstand des jeweils am Signalvolumen anteilsmäßig dominierenden Atompaar abweicht. Dies wurde durch eine weitgehende Ausschaltung der Spindiffusion aufgrund der Wahl einer sehr kleinen Mischzeit (0,03 s) während der Simulation der Spektren erreicht. Spindiffusion bewirkt in experimentellen NOESY Spektren eine zusätzliche Vergrößerung des Signalvolumens und führt somit zu Fehlern bei der Abstandsberechnung. Diese haben wiederum zur Folge, dass weniger mehrdeutigen Signalen eine Zuordnung mit einer hohen Wahrscheinlichkeit zugewiesen werden können, bzw. als eindeutig zugeordnet definiert werden können. Darauf wird später noch näher eingegangen werden.

Die Berechnung der Abstände aus den Signalvolumina erfolgte mit Hilfe eines Kalibrierungsfaktors über die ISPA-Methode. Aufgrund dieser relativ ungenauen Methode, wurde im Durchschnitt eine Abweichung des jeweils ermittelten Abstandes bei richtig zugeordneten zwei- und dreideutigen NOESY-Signalen von etwa 8-10% vom wirklichen Abstand des dominierenden Atompaares innerhalb der Struktur beobachtet. Zur dieser Abweichung trug zusätzlich die Aufsummierung von zueinander nahe im Spektrum liegender Signale bei.

Es wurde somit ein Mittelweg beschritten, da aufgrund des vorhandenen durchschnittlichen Abstandsfehlers eine gewisse Realitätsnähe der Testbedingungen geschaffen wurde, und zum anderen, wegen der geringen Größe dieses Fehlers, es dennoch möglich war, die Zuordnungskapazität der neuen Datenbank auszuloten. Der Großteil der Testreihen wurde deshalb unter diesen eben beschriebenen Bedingungen durchgeführt.

Da bei experimentellen NOESY-Spektren noch weitere Faktoren wie z.B. Spindiffusion, Artefakte, Fehler bei der Volumenintegration oder Unvollständigkeit der sequentiellen Zuordnung einfließen können, ist in der Praxis generell mit größeren Fehlern bei der Abstandsberechnung zu rechnen. Deshalb ist zu erwarten, dass die automatische Zuordnung von experimentellen NOESY-Signalen im allgemeinen nicht so gut ausfallen wird wie bei simulierten Signalen. Aufgrund dieser Tatsachen wurden auch Tests für nicht optimale Fälle durchgeführt, bei denen der jeweils berechnete Abstand sehr ungenau mit dem wirklichen Abstand des signaldominierenden Atompaares innerhalb der Struktur übereinstimmt.

Man muss allerdings berücksichtigen, dass im Laufe der ständigen Optimierung der Verfahren die Abstandbestimmung aus NOESY-Signalen immer genauer wird. Diese wird z.B. seit neueren im AUREMOL nicht mehr mittels eines Kalibrierungsfaktors durchgeführt, sondern unter Anwendung eines fortschrittlicheren Verfahrens über das Programmmodul REFINE.

5.1.2 Unterschiedliche Bedingungen bei verschiedenen Suchradien

Die hier erstellten Testreihen wurden unter verschiedenen eingestellten Suchradien durchgeführt. Es hat sich gezeigt, dass die Anzahl sowie die Eigenschaften der über die chemischen Verschiebungen gefundenen Zuordnungsmöglichkeiten für die in Frage stehenden NOESY-Signale stark vom jeweils eingestellten Suchradius abhingen. Wie zu erwarten, nahm die Anzahl der auf Basis chemischer Verschiebungen gefundener Zuordnungsmöglichkeiten der NOESY-Signale eines Spektrums mit kleiner werdenden Abstandsbeschränkung (=Suchradius) relativ stark ab. So ist die Anzahl von zwei- und dreideutigen NOESY-Signalen bei einen Suchradius von 10 nm im Vergleich zu den eindeutig zugeordneten Signalen im Durchschnitt etwa 4-5 mal so groß. Bei einem Suchradius von 0,6 nm hingegen, beträgt sie nur noch etwas mehr als ein Drittel im Vergleich zu den eindeutigen Signalen. Zudem hat sich gezeigt, dass bei kleineren Suchradius der Anteil der für Abstandsbestimmung interessanten Signale unter den vorhandenen zwei- und dreideutigen Signalen erheblich sinkt. Hierbei handelt es sich um Signale, bei denen eine der gefundenen Zuordnungsmöglichkeiten den Großteil (mind. 90%) des Gesamtvolumens erklärt.

Aufgrund dieser Ausgangslage kann man allgemein sagen, dass die im Rahmen der Arbeit erreichten Verbesserungen bei der Zuordnung zwei- und dreideutiger NOESY-Signale besonders zu Anfang bis etwa zur Mitte des iterativen Strukturbestimmungsprozesses besonders stark auswirken werden.

Ein wesentlicher Unterschied zwischen den hier vorherrschenden Versuchsbedingungen und der Situation in der Praxis bestand darin, dass während der automatisch durchgeführten Zuordnungen von NOESY-NMR-Spektren immer die richtige Struktur als Modellstruktur benutzt wurde. Dies ist in der Praxis nicht der Fall, da ja die richtige Struktur erst ermittelt werden muss und die benutze Modellstruktur in der Regel zu Anfang noch sehr wenig mit der wirklichen Struktur übereinstimmt. In der Praxis kann es sich bei der benutzten

Modellstruktur z.B. zuerst nur um einen ausgestreckten Peptidstrang oder einer Zufallstruktur handeln. Man kann allerdings sagen, dass für sehr große bzw. sehr kleine eingestellte Suchradien die Versuchsbedingungen mit den Bedingungen in der Praxis vergleichbar sind.
Es hat sich nämlich gezeigt, dass bei sehr großen eingestellten Suchradien wie z.B. 10 nm die Modellstruktur als Zuordnungsfilter noch keine Rolle spielt. Sehr kleine Suchradien werden in der Regel vom Benutzer dann gewählt, wenn der Strukturbestimmungsprozess schon weit fortgeschritten ist bzw. die vorhandene Modellstruktur mit der wirklichen Struktur schon relativ gut übereinstimmt. Bei mittelgroßen Suchradien bzw. in einem Stadium in dem die vorhandene Modellstruktur normalerweise von der wirklichen in Frage stehenden Proteinstruktur noch relativ stark abweicht besteht die Gefahr, dass die richtige Zuordnungsmöglichkeit eines mehrdeutigen NOESY-Signals aufgrund der Abstandsbeschränkung ausgeschlossen wird. Es ist somit zu erwarten, dass für mittlere eingestellte Suchradien bei der Zuordnung experimenteller NMR-NOESY-Spektren, im Vergleich zu den hier vorherrschenden Bedingungen, mehr falsche Zuordnungen auftreten werden.

5.2 Verbesserung der Zuordnungsqualität

5.2.1 Gesamtzunahme von Signalzuordnungen

Ein wesentliches Ziel bei der Zuordnung von NOESY-NMR-Spektren ist, möglichst viele der vorhandenen Signale einem bestimmten Atompaar innerhalb der zu untersuchenden Proteinstruktur zuzuordnen. Je mehr Signale zugeordnet werden können, desto mehr Abstandsbeschränkungen können in die Strukturrechnung einfließen und umso schneller und genauer kann man die in Frage stehende Struktur bestimmen. Ein Kernziel der Arbeit war, durch Einsatz einer neuen Datenbank aus Wahrscheinlichkeitsdichteverteilungen die Anzahl zugeordneter zwei- und dreideutiger NOESY-Signale zu erhöhen.
Die Ergebnisse haben gezeigt, dass bei Verwendung der neuen Wahrscheinlichkeitsdichteverteilungen und bei sonst gleichen Bedingungen im Durchschnitt etwa doppelt so viele von der eben genannten Gruppe von NOESY-Signalen zugeordnet werden konnten als im Vergleich zu den bisherigen Verteilungen (Kap 4.2.2.1).
Hierbei muss aber beachtet werden, dass die Häufigkeit von zwei- und dreideutigen NOESY-Signalen, im Vergleich zu den eindeutig zugeordneten Signalen, mit abnehmenden Suchradius und damit mit fortschreitender Verbesserung der zu untersuchenden Proteinstruktur immer weiter abnimmt (Kap. 4.2.1.2). Dies hat zur Folge, dass der prozentuale

Gesamtzuwachs an zugeordneten NOESY-Signalen aufgrund der neuen Verteilungen zu Anfang der Strukturbestimmung am größten ist bzw. gegen Ende am kleinsten ausfällt (s. Abb.5.1).

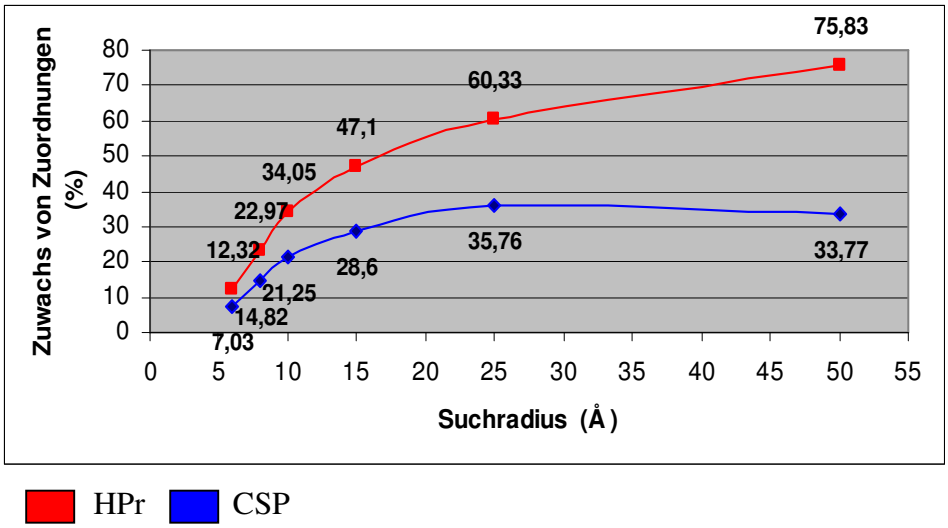


Abbildung 5.1: Prozentualer Anstieg der Anzahl zugeordneter NOESY-Signale. In der Grafik ist der prozentuale Anstieg der Anzahl aller zugeordneten NOESY-Signale du aufgrund der Benutzung von der *Datenbank 3* zu sehen. Die Werte beziehen sich auf die Summe (=100%) aller zugeordneten eindeutigen, zwei -und dreideutigen NOESY-Signale unter Benutzung der früheren Datenbank aus Wahrscheinlichkeitsverteilungen. Die Ergebnisse zeigen die Werte, die für die simulierten 2D-NOESY-NMR-Spektren der Proteine HPr (rote Kurve) und *Tm*CSP (blaue Kurve) bei unterschiedlichen eingestellten Werten für den Suchradius erzielt wurden. Die eingestellte Wahrscheinlichkeitsgrenze betrug P=0.98.

5.2.2 Zunahme von Zuordnungen für unterschiedliche Arten von NOESY-Signalen

Generell werden NOESY-Signale in *intraresiuduale, sequentielle, mittelreichweitige* und *langreichweitige* Signale unterteilt. Die jeweilige Einteilung hängt vom sequentiellen Abstand des hauptsächlich am Signalvolumen beteiligten Atompaares ab. Abstände, welche man aus *langreichweitigen* NOESY-Signalen gewinnt, geben Aufschlüsse über die Tertiärstruktur des zu untersuchenden Proteins. Die anderen Gruppen hingegen, sind vor allem für die Aufklärung von Sekundärstrukturelementen und der Feinstruktur von wichtiger Bedeutung. Generell hat sich gezeigt, das die genannten Gruppen unter den vorhandenen zwei- und dreideutigen NOESY-Signalen mit Hilfe des hier angewandten statistischen bzw. wissensbasierten Verfahrens unterschiedlich häufig zugeordnet werden. Dieser Effekt ist unabhängig von der jeweils benutzten Datenbank. Wie in Kapitel 4.2.2.2 gezeigt, ist der Anteil von zugeordneten *intraresidualen*, *sequentiellen* zwei- und dreideutigen NOESY-Signalen gegenüber *mittel-* und *langreichweitigen* Signalen meist um ein vielfaches größer.

Für *intraresiduale*, *sequentielle* und auch teilweise für *mittelreichweitige* zwei- und dreideutige NOESY-Signale konnte eine deutliche Steigerung an Zuordnungen erzielt werden. Für *langreichweitige* zwei- und dreideutige NOESY-Signale konnte entweder keine oder nur eine sehr geringe Anzahl zusätzlicher Zuordnungen erreicht werden. Auf die genauen Gründe dieser Ergebnisse wird später noch näher eingegangen werden.
Aufgrund der genannten Ergebnisse kann man sagen, dass man durch Einsatz der neuen *Datenbank 3* vor allem einen deutlichen Gewinn an Information über interatomare Abstände von sequentiell relativ nahe gelegenen Atompaaren erhält. Dies wirkt sich insbesondere vorteilhaft auf die korrekte Ausbildung von Sekundärstrukturen sowie der Feinstruktur des zu untersuchenden Proteins während der Strukturrechung aus.

5.2.3 Zunahme der Zuordnungssicherheit

5.2.3.1 Minimierung unerwünschter und falscher Zuordnungen

Das Programm *KNOWNOE* wendet bei der Zuordnung von zwei- und dreideutigen NOESY-Signalen einen statistischen bzw. wissensbasierten Ansatz an. Ziel hierbei ist es, möglichst vielen NOESY-Signalen das Atompaar mit jeweils einer hohen Wahrscheinlichkeit von den vorhandenen Zuordnungsmöglichkeiten zuzuweisen, welches das in Frage stehende Signalvolumen zu mindestens 90 Prozent erklärt. Diese Zuordnungen bilden eine besonders gute Basis für eine exakte Abstandbestimmung. Da es sich hierbei allerdings um ein statistisches Verfahren handelt, kommen zu einem gewissen Prozentsatz auch weniger geeignete Signalzuordnungen vor, welche, zumindest theoretisch, zu einem mehr oder weniger großen Abstandsfehler führen. Diese wiederum können während der Strukturrechnung zur Verzerrung der zu untersuchenden Proteinstruktur führen bzw. die Bestimmung der wirklichen Konformation erschweren. Deshalb war ein wichtiges Ziel der Arbeit, neben der Steigerung der Gesamtzahl von Signalzuordnungen, den Anteil nicht geeigneter Zuordnungen zu minimieren. Generell kann man zwischen zwei Arten von ungeeigneten Zuordnungen unterscheiden:

1. Dem in Frage stehenden NOESY-Signal wird ein Atompaar zugewiesen, das nicht wirklich den größten Teil des vorhandenen Signalvolumens erklärt (hier auch als falsche Zuordnung bezeichnet). Hierbei ist immer eine Zuordnung vorhanden, welche einen größeren Anteil am Signalvolumen, als die jeweils gewählte Zuordnung besitzt.

2. Es erfolgt eine Zuordnung obwohl es kein Atompaar innerhalb der Struktur gibt, welches das betreffende NOESY-Signal zu mindestens 90 % erklärt (hier auch als unerwünschte Zuordnung bezeichnet).

Hierbei ist allerdings zu berücksichtigen, dass, unter Anwendung der ISPA-Methode, der resultierende Abstandsfehler auch bei prozentual geringeren Beiträgen von der gewählten Zuordnungsmöglichkeit zum Signalvolumen, oft relativ gering ausfällt. Wie in Abbildung 5.2 zu sehen führt beispielsweise die Abstandsberechung für ein Atompaar, welches nur 40% eines in Frage stehenden Signalvolumens erklärt, zu einem nur etwa 16% kürzeren Abstand, im Vergleich zu dem Abstand den man erhalten hätte, wenn das Atompaar mehr als 90% des Signalvolumens erklären würde. Aus diesem Grunde kann man sagen, dass auch Zuordnungen bzw. Atompaare mit Beiträgen ab etwa 40% zum Gesamtvolumen eines NOESY-Signals für die Praxis noch brauchbare Abstände liefern.

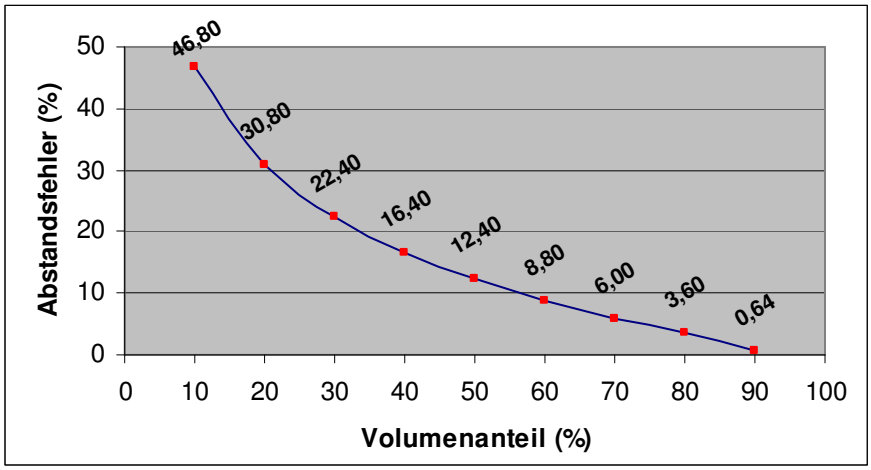


Abbildung 5.2: Abstandsfehler. Die Grafik zeigt den prozentualen Abstandsfehler, der sich theoretisch ergeben würde, wenn einem NOESY-Signal ein Atompaar zugeordnet wird, welches sein Signalvolumen in Wirklichkeit zu 10%, 20% usw. erklärt. Die Überlegung beruht auf der Annahme, dass zwischen dem Abstand r zweier Atome und dem Signalvolumen V eines NOESY-Signals die Beziehung $V \sim 1/r^6$ gilt (ISPA-Methode).

Es konnte gezeigt werden, dass die prozentualen Anteile dieser Art von Zuordnungen unter Benutzung der neuen *Datenbank 3* deutlich erhöht werden konnte. Dies zeigt sich insbesondere bei sehr kleinen eingestellten Suchradien (s. Abb. 5.3), da bei diesen, wie bereits erwähnt, die Anteile von Atompaaren mit relativ geringen Beiträgen zum Signalvolumen unter den potentiellen Zuordnungsmöglichkeiten, sehr hoch ist.

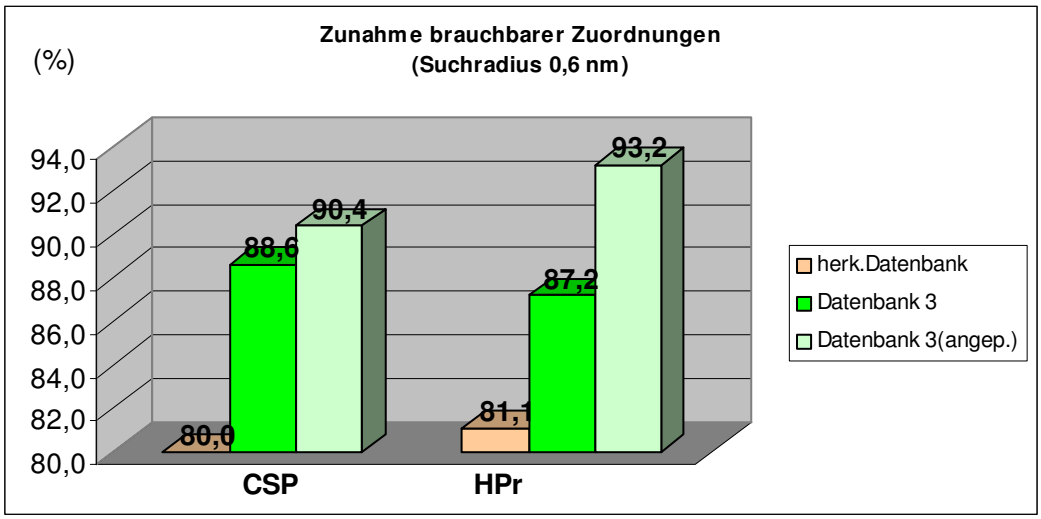


Abbildung 5.3: Zunahme brauchbarer Zuordnungen. Die Grafik zeigt jeweils die prozentualen Anteile von zugeordneten zwei- und dreideutigen NOESY-Signalen, deren zugewiesene Zuordnung mindestens 40% vom Signalvolumen erklärt (=brauchbare Zuordnung). Der orangefarbene Balken zeigt die Werte bei der Benutzung der herkömmlichen Datenbank, der grüne Balken steht für die neue *Datenbank 3* und der hellgrüne Balken für die neue *Datenbank 3* unter Berücksichtigung vom Suchradius. Die Werte sind für die simulierten 2D-NOESY-NMR-Spektren der Proteine CSP und HPr bei einem eingestellten Suchradius von 0,6nm dargestellt. Die Wahrscheinlichkeitsgrenze betrug P=0,98.

Die Auswirkung von falschen Zuordnungen für die Abstandsbestimmung muss, in Anbetracht der oben gemachten Überlegungen, differenzierter betrachtet werden. Anhand der Abbildung 5.2 kann man sagen, dass falsche Zuordnungen zu einem umso größeren Abstandsfehler (zu kurzer Abstand) führen werden je größer der Anteil der richtigen Zuordnung am Gesamtvolumen des in Frage stehenden NOESY-Signals ist. Somit haben falsche Zuordnungen von NOESY-Signalen, die von keinem Atompaar zu mindestens 90% erklärt werden generell eine geringere negative Auswirkun auf die Abstandsbestimmung als falsche Zuordnungen innerhalb von NOESY-Signalen, welche durch ein bestimmtes Atompaar stark (mind. 90%) dominiert werden. Es hat sich ergeben, dass unter Anwendung der neuen *Datenbank 3*, im Vergleich zur herkömmlichen Datenbank, innerhalb beider genannter Gruppen, der prozentuale Anteil an falschen Zuordnungen deutlich verringert werden konnte. Dies zeigte sich besonders deutlich bei sehr kleinen eingestellten Suchradien unter denen, wie bereits in Kapitel 4.2.2.4 gezeigt, der prozentuale Anteil falscher Zuordnungen im allgemeinem besonders hoch ist.

Es konnte gezeigt werden, dass unter Anwendung der neuen *Datenbank 3* und bei jeweils denselben Versuchbedingungen, die Fehlerquoten wesentlich geringer ausfielen. So konnte bei einem Suchradius von 0,6 nm und einer eingestellten Wahrscheinlichkeitsgrenze von P=0,98 für die Gruppe von NOESY-Signalen, welche durch kein bestimmtes Atompaar

dominiert (mind. 90%) werden, die Fehlerquote von 39% auf 28% (Protein CSP) und von 37% auf 31% (Protein HPr) erniedrigt werden. Für die entsprechend andere Gruppe konnte die Fehlerquote von durchschnittlich 14-15% auf etwa 3-4% verringert werden. Für die Strukturbestimmung ist die erreichte Reduzierung von falschen Zuordnungen innerhalb der letztgenannten Gruppe von NOESY-Signalen besonders wichtig, da hier falsche Zuordnungen zu besonders großen Abstandsfehlern führen. Darüber hinaus wäre hier noch anzumerken dass, wie in Kapitel 4.2.2.8 gezeigt, die Anteile falscher Zuordnungen mit steigendender Wahrscheinlichkeitsgrenze, unabhängig von der benutzten Datenbank, sinken. Allerdings geschieht dies unter Verwendung der neuen *Datenbank 3* schneller.

Besonders destruktiv auf die Strukturbestimmung wirken sich solche Zuordnungen von NOESY-Signalen aus, die fälschlicherweise eine räumliche Nähe von sequentiell weit auseinanderliegenden (langreichweitigen) Atompaaren propagieren. Während der Strukturrechung genügt bereits nur eine dieser Zuordnungen, um die Struktur in eine völlig falsche Konformation zu führen. Solche Zuordnungen können besonders zu Anfang der Strukturbestimmung verhängnisvoll sein, da in diesem Moment eine einigermaßen richtige Grobfaltung des zu untersuchenden Proteins für den Erfolg des weiteren Strukturbestimmungsprozesses entscheidend ist. Zusätzlich ist anzumerken, dass sich falsche Zuordnungen, aufgrund des zu Anfang vorliegenden relativ unrealistischen Strukturmodells, nur sehr schwer erkennen bzw. ausschließen lassen. Deshalb ist es wichtig, dass es gar nicht erst zu solcher Art von Zuordnungen kommt. Die Testreihen zeigten, dass unter Benutzung der herkömmlichen Datenbank oft solche gefährlichen Zuordnungen vorkommen. So konnten z.B. bei einem Suchradius von 10,0 nm und einer gewählten Wahrscheinlichkeitsgrenze von P=0,98 für das simulierte 2D-NOESY-NMR Spektrum vom Protein CSP gleich drei falsch zugeordnete NOESY-Signale mit Zuordnungen bzw. Atompaaren mit jeweils über 20 Aminosäuren Sequenzabstand beobachtet werden. Bei Benutzung der neuen *Datenbank 3* wurde hingegen, unter der eingestellten Wahrscheinlichkeitsgrenze von P=0,98, unabhängig von Spektrum oder Suchradius, keine falsche langreichweitige Zuordnung beobachtet. Dieses Resultat lässt sich, wie aus den Grafiken der Kapitel 4.1.2.3 deutlich zu entnehmen, mit der allgemein geringen statistischen Wahrscheinlichkeit für eine räumlichen Nähe ($\leq$ 0,5nm) sequentiell weit entferner Atome begründen. Aus demselben Grund erhielten, wie in Kapitel 4.2.2.6 gezeigt, unter Benutzung der neuen *Datenbank 3,* langreichweitige zwei- und dreideutige NOESY-Signale, im Falle einer erstellten Zuordnung, immer jeweils eine falsche und zugleich sequentiell kürzere Zuordnungsmöglichkeit (=Atompaar) zugewiesen. Unter Benutzung der herkömmlichen Verteilungen konnten hingegen wenige richtige

langreichweitige Zuordnungen beobachtet werden. Da diese, genau wie die bereits erwähnten falschen langreichweitigen Zuordnungen, vermutlich auf Unstimmigkeiten innerhalb der Verteilungen zurückzuführen sind, kann man solche Zuordnungen nicht positiv bewerten.
Zusammenfassend kann man aufgrund der Ergebnisse sagen, dass die neuen Wahrscheinlichkeitsdichteverteilungen die statistischen Verteilungen interatomarer Abstände innerhalb von Proteinen, im Vergleich zu den herkömmlichen Verteilungen, besser wiedergeben. Die dadurch erreichte größere Sicherheit der Zuordnungen führte wiederum zu geringeren Abstandsfehlern, wie bereits in Kapitel 4.2.2.7 gezeigt werden konnte.

5.2.3.2 Stabilität gegenüber falschen Abständen

Generell muss man bei experimentellen NOESY-NMR-Spektren gegenüber simulierten NOESY–NMR-Spektren, aufgrund der potentiell größeren Anzahl von Fehlerquellen, mit einer größeren Anzahl bei der Abstandsbestimmung rechnen. Da die hier angewandte statistische Zuordnungsmethode den aus den Signalvolumen ermittelten Abstand als Grundlage zur Berechnung der wahrscheinlichsten Zuordnung benutzt, ist für die Stabilität des Verfahrens eine gewisse Toleranz gegenüber Abstandfehlern besonders wichtig.
Das heißt, bezüglich der Anwendung, dass sich Abstandsfehler möglichst gering auf Änderungen der Zuordnungsanzahl wie auch die zahlenmäßigen Verhältnisse zwischen richtigen und falschen Zuordnungen auswirken sollten.
Bei den hier angewandten statistischen Verfahren ist, aufgrund von Abstandsfehlern, generell mit einer Verringerung von Zuordnungen zu rechnen. Das lässt sich damit begründen, dass Abstandsfehler in allgemeinen zu einem statistisch unwahrscheinlicheren Atomabstand führen. Dies hat wiederum zur Folge, dass dadurch öfter für keine der jeweils in Frage stehenden Zuordnungsmöglichkeiten für ein bestimmtes NOESY-Signal eine hohe Wahrscheinlichkeit berechnet werden kann. Das Verhältnis zwischen richtigen und falschen Zuordnungen sollte hingegen weitgehend unverändert bleiben.
Die Versuche in Kapitel 4.2.2.10 anhand von künstlich eingebauten Abstandsfehlern haben gezeigt, dass, im Gegensatz zur herkömmlichen Datenbank und unter Verwendung der neuen *Datenbank 3*, das Resultat der automatisch erstellten NOESY-Signalzuordnungen weitgehend den hier gemachten theoretischen Überlegungen entspricht. Als Ausnahme ist hierbei die leichte Zunahme von Zuordnungen beim 2D-NOESY-Spektrum vom Protein CSP bei künstlich verlängerten Abständen zu nennen (s. Kap. 4.2.2.10). Der Grund hierfür liegt vermutlich in der relativ geringen Größe des Proteins im Vergleich zu den meisten anderen Proteinen innerhalb der benutzen Strukturdatenbasis. Da innerhalb größerer Proteine

entsprechend größere Abstände statistisch wahrscheinlicher sind, konnte man hier eine erhöhte Anzahl von Zuordnungen beobachten.
Ein besonders wichtiges Ergebnis war, dass, im Vergleich zur Benutzung der herkömmlichen Datenbank, das Verhältnis zwischen richtigen und falschen Zuordnungen, auch durch die künstlich eingebauten Abstandsfehler, wesentlich geringeren Schwankungen unterlag. Unter Benutzung der herkömmlichen Datenbank fiel auf, dass zu kurze Abstände zu einer sehr starken Zunahme der Fehlerquote bzw. zu lange Abstände zu einer unverhältnismäßig starken Abnahme an Zuordnungen führten.
Anhand der Ergebnisse kann man zusammenfassend festhalten, dass bei Verwendung der neuen *Datenbank 3* eine deutlich höhere Zuverlässigkeit und Stabilität für die automatische Zuordnung von experimentellen NOESY-NMR- Spektren zu erwarten ist.

5.2.4 Die Bedeutung der spezifischen Eigenschaften der Datenbanken für die Zuordnungsqualität

Hier soll erläutert werden, welche Faktoren für die Qualität der Zuordnung von zwei- und dreideutigen NOESY-Signalen sich als besonders wichtig erwiesen haben. Im Zentrum der Betrachtung stehen hier vor allem die Eigenschaften der jeweils eingesetzten Datenbanken bzw. ihrer Verteilungen. Dabei soll insbesondere diskutiert werden, welche Unterschiede und Erweiterungen der neuen Datenbanken, verglichen mit der herkömmlichen Datenbank, zur Verbesserung der erreichten Zuordnungsqualität der NOESY Spektren beitrugen bzw. eher redundant waren.

5.2.4.1 Erweiterung der Abstandsklassen

Insgesamt wurde die Anzahl der Wahrscheinlichkeitsverteilungen von 1593 bei der früheren Datenbank auf 220280 (*Datenbank 3)* erhöht. Die *Datenbank 1* beinhaltet 3620 Verteilungen und die *Datenbank 2* 16483 Verteilungen. Die Erhöhung der Anzahl an Wahrscheinlichkeitsverteilungen ist eine unmittelbare Folge der hier durchgeführten Erweiterung von Abstandsklassen. Das heißt, es wurden wesentlich mehr unterschiedliche Gruppen von Atompaaren definiert aus dessen Abständen (=Abstandklasse) eine bestimmte Wahrscheinlichkeitsdichteverteilung generiert wurde. Ziel der Erweiterung war im wesentlichem, für möglichst viele unterschiedliche Zuordnungsmöglichkeiten eine repräsentative Verteilungskurve zu Verfügung zu stellen. Dies ermöglicht wiederum für eine größere Anzahl zwei- und dreideutiger NOESY-Signale die wahrscheinlichste

Zuordnungsmöglichkeit zu finden. Für die Bildung einer bestimmten Abstandklasse bzw. Definition einer bestimmten Gruppe von Atompaaren wurden folgende Kriterien angewandt:

1. Lageposition der Atome eines Paares innerhalb ihrer Aminosäure (z.B. HA).
2. Relativer Sequenzabstand beider Atome (für jeweils 0,1,2..8 Aminosäuren).
3. Zugehörigkeit der Atome zu einem bestimmten Aminosäuretyp.

Hierbei handelt es sich vor allem um Kriterien, die sich auf den räumlichen Abstand zweier Atome innerhalb eines Proteins auswirken können. Im folgendem soll auf die Bedeutung der Berücksichtigung dieser genannten Kriterien bei der Bildung von Abstandsklassen für die Zuordnungsqualität näher eingegangen werden.
Bei den Verteilungen der *Datenbank 3* wurden alle aufgeführten Kriterien zugleich angewandt. Bei den Verteilungen der *Datenbank 1* bzw. *Datenbank 2* wurde zu Testzwecken entweder das erste oder das dritte Kriterium nicht berücksichtigt. Demzufolge sind, im Fall der der *Datenbank 1,* Abstände von Atomen mit unterschiedlichen Lagepositionen innerhalb der Aminosäure und, im Fall der *Datenbank 2,* Abstände von Atomen aus jeweils unterschiedlichen Aminosäuren in eine Verteilung integriert worden.
Unter Benutzung der *Datenbank 3,* welche die meisten Verteilungen enthält, konnte die Anzahl zugeordneter zwei- und dreideutiger NOESY-Signale, im Vergleich zur früheren Datenbank, in etwa verdoppelt werden. Bei Benutzung der *Datenbank 2* und *1*, war die erreichte Steigerung gegenüber der früheren Datenbank deutlich geringer. Dies zeigt, dass die gleichzeitige Unterscheidung von Atomen nach Lageposition und Aminosäurezugehörigkeit bei der Bildung von Abstandklassen zu einem zusätzlichen Gewinn statistisch relevanter Abstandsinformation führt. Dies kann man auch daran sehen, dass, wie bereits erwähnt, die Kurvenverläufe verschiedener Atompaare mit gleichem relativem sequentiellen Abstand sich oft sehr stark unterscheiden können.
Allerdings ist zu beachten, dass die hier erreichte Verdopplung an zugeordneten NOESY-Signalen unter Einsatz einer um das 138 fache höheren Anzahl an Verteilungskurven bewerkstelligt wurde. Bei der hier eingestellten Wahrscheinlichkeitsgrenze von P=0,98 konnten, trotz der hohen Anzahl an vorhandenen Verteilungen, immer noch etwa die Hälfte der in Frage stehenden zwei- und dreideutigen NOESY-Signale nicht zugeordnet werden. Anhand dieser genannten Aspekte lässt sich schließen, dass sich viele der in der *Datenbank 3* vorhandenen Verteilungen sich relativ wenig voneinander unterscheiden. Wie bereits bekannt, gleichen sich die Kurvenverläufe unterschiedlicher Atompaare vor allem mit zunehmendem

Sequenzabstand (ab etwa drei Aminosäuren) immer stärker aneinander an. Dies machte sich, wie in Kapitel 4.2.2.2 gezeigt, daran bemerkbar, dass deutlich geringere Anteile der vorhandenen mittelreichweitigen und langreichweitigen zwei- und dreideutigen NOESY-Signalen, im Vergleich zu den kurzreichweitigen Signalen, zugeordnet werden konnten.
Begründen lässt sich dies außerdem durch die im allgemeinem schnell abnehmende statistische Wahrscheinlichkeit einer räumlichen Nähe (hier <0,5nm) zweier Atome bei zunehmenden sequentiellen Abstand. Dies wird besonders in den Grafiken der Abbildung 4.8 aus Kapitel 4.1.2.3 deutlich. Bereits bei Atomen mit einem relativen Sequenzabstand von vier Aminosäuren ist der durchschnittliche räumliche Abstand mit etwa 1,0 nm schon doppelt so groß wie der für NOESY-Signale detektierbare Abstandsbereich. Wie die Ergebnisse aus Kapitel 4.2.2.11 zeigen, führte die Bildung von Abstandsklassen für Atome mit jeweils gleichem sequentiellem Abstand bis maximal 4 Aminosäuren zu einer Steigerung von Zuordnungen. Durch die im Rahmen der Arbeit durchgeführte Erweiterung der Abstandsklassen für Atome mit sequentiellen Abständen von 5,6,7 und 8 Aminosäuren, konnte hingegen keine weitere Steigerung an Zuordnungen erzielt werden. Dies war, wie bereits gesagt, aufgrund der geringen Unterschiede dieser Verteilungen untereinander, zu erwarten.
Für mittelreichweitige NOESY-Signale wurden dennoch mehr Zuordnungen erwartet, da innerhalb von Sekundärstrukturen sich Atome mit mittelreichweitigen und sequentiellen Abständen oft regelmäßig räumlich sehr nahe kommen. Vermutlich führte die gegenseitige Überlagerung von Abständen aus unterschiedlichen Sekundärstrukturelementen, Schleifen-regionen sowie anderen weniger geordneten Bereichen der Proteine zu einer allgemeinen Verringerung der Wahrscheinlichkeitsdichten für kürzere Abstände. Es wäre deshalb von Vorteil gewesen gesonderte Abstandsklassen bzw. Verteilungen aus Abständen aus jeweils unterschiedlichen Sekundärstrukturelementen zu erzeugen.
Die Tatsache, dass nicht alle der vorhandenen kurzreichweitigen zwei- und dreideutigen NOESY-Signale zugeordnet werden konnten, machte deutlich, dass es auch viele unterschiedliche Atome mit kurzen sequentiellen Abständen gibt, welche eine ähnliche Wahrscheinlichkeitsdichteverteilung aufweisen. Dies trifft insbesondere bei Wasserstoff-atompaaren des Proteinrückrats zu, bei denen ihr jeweiliger Abstand in der Regel kaum von der Aminosäurezugehörigkeit abhängt.
Generell kann man sagen, dass sich, aufgrund der Vielzahl untereinander ähnlicher Verteilungen, noch ein hoher Anteil redundanter Information innerhalb der Datenbank befindet. Es wäre deshalb sinnvoll Gruppen von Wahrscheinlichkeitsdichteverteilungen mit

ähnlichem Kurvenprofil durch eine Verteilung zu ersetzen. Dies würde im weiterem zu einer erheblichen Reduktion der Datenmenge führen.

Zusammenfassend kann man sagen, dass die hier durchgeführte erweiterte Differenzierung von Atomen bei der Bildung von Abstandsklassen, bezüglich ihrer Aminosäureherkunft und Lageposition innerhalb der Aminosäure, zu einem zusätzlichen Gewinn struktureller Information führte. Diese machte sich in einer etwa doppelten so großen Anzahl zugeordneter zwei- und dreideutiger NOESY-Signale bemerkbar. Die erweiterte Differenzierung lohnte sich allerdings nur für Atome mit relativen Sequenzabständen von 0,1,2,3 und 4 Aminosäuren. Bei Atomen mit jeweils sequentiellen Abständen von mehr als 4 Aminosäuren hat sich hingegen eine erweiterte Differenzierung weder nach Sequenzabstand oder Aminosäureherkunft sowie nach Lageposition innerhalb der Aminosäure für die Anwendung als lohnend erwiesen

5.2.4.2 Erhöhung der Datenauflösung

Ein wesentlicher Unterschied zwischen den neuen Wahrscheinlichkeitsdichteverteilungen gegenüber den früheren Wahrscheinlichkeitsverteilungen besteht in der um das hundertfache erhöhten Datenauflösung von jeweils 10000 Datenpunkten. Damit sollte erreicht werden, alle eventuell vorhandenen Strukturen einer Verteilungskurve zwischen zwei Atomen sichtbar zu machen und somit entsprechend viel statistisch relevante Abstandsinformation zu gewinnen.

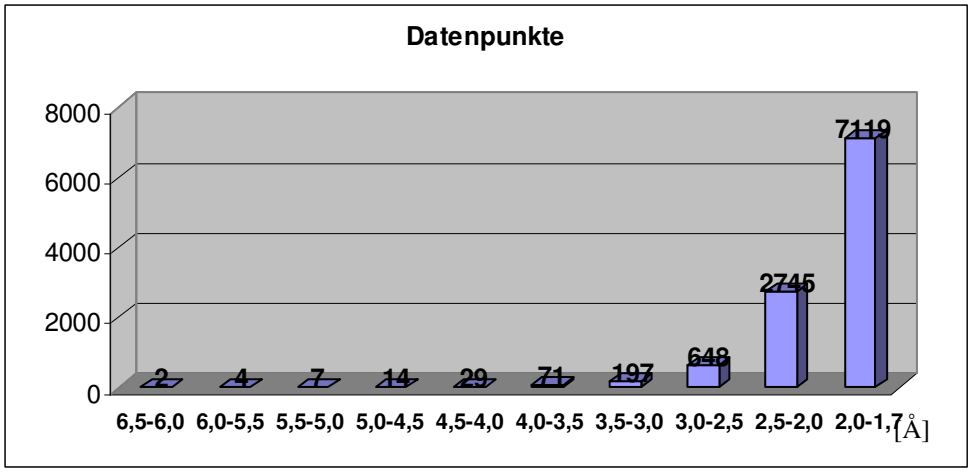


Abbildung 5.4: Verteilung von Datenpunkten. Das Diagramm zeigt, in wie viele Datenpunkte die erzeugten Volumenwahrscheinlichkeitsdichteverteilungen bezüglich unterschiedlicher äquidistanter Abstandsintervalle aufgelöst sind. So entspricht z.B. der Abstandsbereich r_1 bis r_2 mit r_1=3,0 Å und r_2=3,5Å, nach Umrechung mit $V=(1/r^6)x$ $Å^9$, einem Volumenbereich v_1 bis v_2 mit v_1~0,00137 $Å^3$ und v_2~0,00054 $Å^3$. Dieser ist in den Volumenwahrscheinlichkeitsdichteverteilungen, aufgrund der gewählten Schrittweite von 0,0000042 $Å^3$, in etwa 197 Datenpunkte aufgeteilt.

Dass diese hohe Datenauflösung sinnvoll ist konnte man daran sehen dass, wie in Kapitel 4.2.2.9 gezeigt, die Benutzung gleicher Verteilungen, aber mit einer niedrigeren Datenauflösung, zu einer verringerten Anzahl von Zuordnungen führte.
Eine noch höhere Auflösung der Verteilungen wurde, aufgrund der relativ geringen zu erwartenden Steigerung an Zuordnungen, als nicht sinnvoll erachtet. Dies soll anhand der Abbildung 5.4 erläutert werden. Sie zeigt die Anzahl der berechneten Datenpunkte einer Volumenwahrscheinlichkeitsdichteverteilung bezogen auf unterschiedliche äquidistante Abstandsintervalle. Die deutlich erkennbare ungleichmäßige Verteilung beruht darauf, dass äquidistante Abstandsintervalle mit jeweils unterschiedlichen Grenzen, aufgrund der Beziehung: Volumen=1/Abstand^6, unterschiedlich große Volumenbereiche einschließen. Weiter kann man erkennen, dass die meisten der für die NMR interessanten Abstandsintervalle (<0,5nm) bereits in mehr Datenpunkte unterteilt sind als, in Anbetracht der technisch möglichen Messgenauigkeit, für die Abstandsbestimmung sinnvoll ist. Deshalb ist durch eine noch höhere Auflösung der Verteilungen mit keinem weiteren Gewinn an struktureller Information zu rechnen.

5.2.4.3 Rolle der Strukturdatenbasis und des Kurvenglättungs-verfahren

Wie bereits erwähnt, konnte das Verhältnis zwischen richtigen und falschen Zuordnungen durch Einsatz der neuen Datenbanken (1-3), im Vergleich zur früheren Datenbank, deutlich verbessert werden. Beim Vergleich der neuen Datenbanken untereinander zeigten sich, bezüglich der Fehlerquoten, jedoch keine nennenswerten Unterschiede. Daraus lässt sich schließen, dass die hier durchgeführte Erweiterung der Abstandsklassen zwar für eine größere Anzahl an Zuordnungen sorgte, aber nicht für die beobachtete Verbesserung der Zuordnungssicherheit. Es ist deshalb anzunehmen, dass die wesentlich größere zugrundeliegende Proteinstrukturdatenbasis und das angewandte Kurvenglättungsverfahren (Summierung über Gaußkurven) bei der Erzeugung der Verteilungen für die erreichte Verbesserung der Zuordnungssicherheit im wesentlichem verantwortlich waren. Dies hat folgende Gründe:

1. Aufgrund der größeren Anzahl von Proteinen konnten entsprechend mehr Abstände bzw. Volumina in eine Verteilung integriert werden.

2. Die relativ geringe sequentielle Übereinstimmung der Proteine innerhalb der Datenbasis von weniger als 25% sorgte für eine möglichst hohe Repräsentativität der Abstände bezüglich aller vorkommenden Proteine.

3. Das hier angewandte Kurvenglättungsverfahren (Summierung über Gaußkurven) sorgte für einen gleichmäßigen Kurvenverlauf. Dies wird durch die Fähigkeit des Verfahrens zur Unterdrückung s.g. statistischer Ausreißer, wie auch der harmonischen Weiterführung der Kurve innerhalb statistisch schlecht definierter Bereiche, erreicht.

Aus den genannten Fakten kann man schließen, dass die neuen Verteilungen, im Vergleich zu den früheren Verteilungen, repräsentativer sind, was somit die erreichte Verbesserung der Zuordnungssicherheit erklärt.

5.2.5 Grenzen der Anwendbarkeit der neuen Wahrscheinlichkeitsdichteverteilungen

Die Grenzen der Anwendbarkeit der neuen Wahrscheinlichkeitsdichteverteilungen für das im Programm *KNOWNOE* angewandte statische Verfahren zur Zuordnung zwei- und dreideutiger NOESY-Signale ist das Thema des folgenden Kapitels.

5.2.5.1 Langreichweitige NOESY-Signale

Wie bereits erwähnt, konnte durch Einsatz der neuen Datenbank die Anzahl der zugeordneten zwei- und dreideutigen NOESY-Signale deutlich erhöht werden. Allerdings kam es insbesondere für langreichweitige NOESY-Signale, trotz Einsatzes großer Mengen struktureller Abstandsinformationen, zu keiner Steigerung der Zuordnungsanzahl. Das Grundproblem liegt, wie bereits erwähnt, daran dass die Wahrscheinlichkeit für eine räumliche Nähe (<0,5nm) zweier sequentiell weit auseinanderliegender Atome, innerhalb eines Proteins, relativ unwahrscheinlich ist. Hinzu kommt, dass die Verteilungskurven von sequentiell weit auseinander liegenden Atomen sich untereinander kaum unterscheiden.

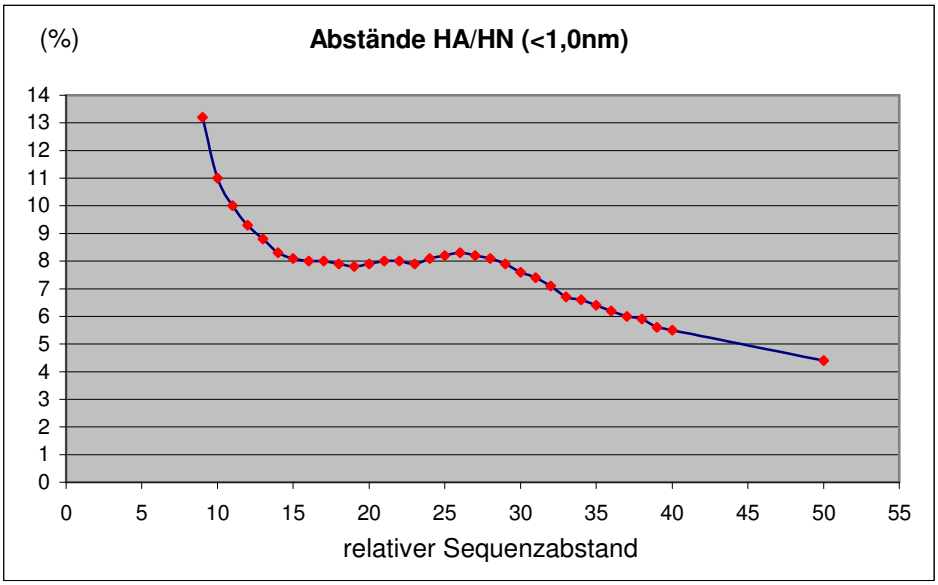


Abbildung 5.5: Anteile kleiner Abstände. Die Grafik zeigt den jeweils prozentualen Anteil von Abständen von jeweils kleiner als 1,0 nm zwischen HA- und HN-Atomen für jeweils unterschiedliche sequentielle Abstände. Die Daten wurden aus 1107 Proteinstrukturen gewonnen.

Es haben sich auch keine Hinweise ergeben, dass es, aufgrund der Proteingrobfaltung, für Atome mit jeweils bestimmten sequentiellen weiten Abständen (z.B. 20 Aminosäuren) zu einer besonders starken Häufung von räumlichen Annäherungen kommt. Dies kann man in Abbildung 5.5 deutlich sehen. Sie zeigt den jeweiligen prozentualen Anteil von Abständen zwischen HA und HN-Atomen mit jeweils weniger als 1,0 nm für unterschiedliche sequentielle Abstände. Wie aus dem Kurvenverlauf zu entnehmen ist, liegen die Werte bereits ab einem sequentiellen Abstand von 9 Aminosäuren mit abfallender Tendenz unter 14 Prozent.

Aufgrund dieser Tatsachen, kann, im Fall von langreichweitigen Zuordnungsmöglichkeiten, für ein NOESY-Signal keine Entscheidung mit einer statistisch hohen Sicherheit (z.B. > 90%) getroffen werden. Generell lag der Anteil von zugeordneten zwei- und dreideutigen langreichweitigen NOESY-Signalen bei dem hier untersuchten Fällen bei unter 22 Prozent. Sowohl unter Einsatz der neuen *Datenbank 3* als auch der früheren Datenbank ist diesen zugeordneten Signalen in über 90% der Fälle (bei der neuen Datenbank praktisch immer) das jeweils falsche Atompaar zugewiesen worden. In diesen Fällen lag immer ein um weniger als vier Aminosäuren auseinander liegendes Atompaar, neben jeweils langreichweitigen Zuordnungsmöglichkeiten, als Zuordnungsmöglichkeit vor. Das Verfahren wählte hierbei grundsätzlich die sequentiell kürzere Zuordnungsmöglichkeit aus, da diese, wie bereits erwähnt, generell statistisch wahrscheinlicher ist. Prinzipiell hat somit das Zuordnungsverfahren keinen Fehler gemacht.

Es muss an dieser Stelle allerdings erwähnt werden, dass nur in etwa 15 % (CSP) bzw. 19 % (HPr) der Fälle bei diesen Zuordnungskonstellationen wirklich eine Zuordnung mit einer hohen Wahrscheinlichkeit (98 %) erstellt wurde (untersucht bei einer Wahrscheinlichkeitsgrenze von P=0,98 und einem Suchradius von 0,6 nm). Das heißt, dass auch meist für die jeweils sequentiell kürzere Zuordnungsmöglichkeit nicht die erforderliche Wahrscheinlichkeit für eine Übernahme in die Liste der eindeutig zugeordneten NOESY-Signale berechnet werden konnte. Dies zeigt, dass das Zuordnungsverfahren auch innerhalb kleiner Abstandsbereiche (<0,6 nm) zwischen wahrscheinlichen und unwahrscheinlichen Abständen bzw. Volumina recht gut unterscheiden kann.
Zusammenfassend kann man sagen, dass sich auch unter Verwendung der neuen Datenbank langreichweitige NOESY-Signale grundsätzlich einer richtigen Zuordnung entziehen.
Eine Möglichkeit solche Signale dennoch zugänglich zu machen wäre beispielsweise die Verwendung einer Datenbasis aus Proteinen, welche mit dem speziell zu untersuchenden Protein einen realtiven hohen Verwandtschaftsgrad aufweisen. Die daraus resultierenden Verteilungen wären wesentlich repräsentativer bezüglich der jeweils in Frage stehenden Zuordnungsmöglichkeiten. Vor allem dürften sich, aufgrund der häufig auftretenden strukturellen Ähnlichkeit verwandter Proteine, auch unterschiedliche Verteilungsverläufe von sequentiell weiter auseinanderliegenden Atomen zeigen. Diese sind vermutlich in den Verteilungen der neuen Datenbank, aufgrund der vielen unterschiedlichen Proteine innerhalb der Datenbasis, herausgemittelt worden. Allerdings muss man hierbei bedenken, dass für jedes neue zu untersuchende Protein eine entsprechend neue Datenbank erzeugt werden müsste. Dies ist, aufgrund der im Rahmen der Arbeit bereitgestellten Programme und Funktionen, zu jeder Zeit und ohne größeren Aufwand möglich.

5.2.5.2 Abhängigkeit der Zuordnungssicherheit vom Suchradius

Wie bereits erwähnt, fielen die Anteile von falsch zugeordneten zwei- und dreideutigen NOESY-Signalen unter Einsatz der neuen Datenbank, im Vergleich zu den früheren Verteilungen, deutlich geringer aus. Das Verhältnis zwischen unerwünschten und erwünschten Zuordnungen konnte auch, wenn nur geringfügig, verbessert werden. Trotzdem blieb der festgestellte Trend der Zunahme der Anteile unerwünschter Zuordnungen, und dadurch verbunden auch eine entsprechende Anhäufung falscher Zuordnungen, mit kleiner werdendem Suchradius bestehen. Im optimalen Fall dürften, aufgrund der Theorie, maximal 1-P (P = eingestellte Wahrscheinlichkeitsgrenze) des Anteils der zugeordneten zwei- und dreideutigen NOESY-Signale eine falsche oder unerwünschte Zuordnung aufweisen. Bei der

eingestellten Wahrscheinlichkeitsgrenze von P= 0,98 wird dieser Wert, also 0,2 bzw. 2%, ab einem Suchradius von etwa 1.5 nm bereits deutlich überschritten (s. Kap. 4.2.2.4). Im folgendem sollen die Gründe für dieses Phänomen näher diskutiert werden.

5.2.5.3 Unerwünschte Zuordnungen

Wie bereits in Kapitel 4.2.2.6 gezeigt, werden besonders häufig solche zwei- und dreideutigen NOESY-Signale falsch zugeordnet, welche durch keine der jeweils gegebenen Zuordnungsmöglichkeiten zum Großteil (hier 90%) erklärt werden. Auf die Gründe hierfür wird noch im nächsten Kapitel näher eingegangen werden. Das Signalvolumen verteilt sich bei der eben genannten Gruppe von Signalen oft mehr oder weniger gleichmäßig auf die jeweiligen Zuordnungsmöglichkeiten auf. Das bedeutet, dass sich die Zuordnungsmöglichkeiten bezüglich ihres Abstandes innerhalb der Struktur entsprechend wenig voneinander unterscheiden. Da, wie bereits bekannt, mit kleiner werdenden Suchradius die Anteile dieser Gruppe unter den zwei- und dreideutigen NOESY-Signalen stark zunimmt, wächst somit die Gefahr eine falsche Zuordnung zu erstellen entsprechend an. Im optimalen Fall werden solche NOESY-Signale gar nicht zugeordnet. Es hat sich allerdings gezeigt, dass sich bei einem Suchradius von 0,6 nm und einer Wahrscheinlichkeitsgrenze von P=0,98 etwa die Hälfte der jeweils zugeordneten zwei – und dreideutigen NOESY-Signale aus der eben genannten Art von Signalen zusammensetzt. Dies ist, wie in Kapitel 4.2.2.4 gezeigt, weitgehend unabhängig von der jeweils benutzen Datenbank. Ein Hauptgrund ist vermutlich darin zu sehen, dass in Abstandsbereichen von wenigen hundertstel Nanometern bzw. den entsprechenden Volumenbereichen im allgemeinem keine großen Änderungen der Wahrscheinlichkeitsdichten innerhalb der Verteilungen zu erwarten sind. Hinzu kommt, dass diese, falls wirklich vorhanden, vermutlich aufgrund von immer gegenwärtigen Messfehlern, bei der Abstandsbestimmung, herausgemittelt wurden. Aufgrund der angenommenen Beziehung $V=ar^{-6}$, führen allerdings bereits relativ kleine Abstandsunterschiede zwischen vergleichbaren Atompaaren zu verhältnismäßig großen Unterschieden im resultierenden Signalvolumen innerhalb des Spektrums. Wenn ein bestimmten Atompaar mit beispielsweise einem Abstand von 0,3 nm zu über 99% ein bestimmten Signalvolumen erklärt, würde ein anderes vergleichbares Atompaar mit einem nur 0,02 nm größeren Abstand das Signal nur noch zu etwa 60% erklären. Aufgrund des geringen Abstandsunterschieds ist, wie bereits angemerkt, mit ähnlich großen resultierenden Wahrscheinlichkeiten für die beiden genannten Atompaare zu rechnen. Das erklärt somit die hohen Anteile von zugeordneten zwei- und

dreideutigen NOESY-Signalen, welche durch kein bestimmtes Atompaar zu mindestens 90% erklärt werden.

5.2.5.4 Falsche Zuordnungen

Wie bereits erwähnt, konnten die prozentualen Anteile falscher Zuordnungen sowohl für zugeordnete zwei- und dreideutige NOESY-Signale, welche durch ein bestimmtes Atompaar zu 90% erklärt werden, als auch für die entsprechend andere Gruppe (=unerwünschte Zuordnungen), deutlich reduziert werden. Trotzdem blieb in der Gruppe der unerwünschten Zuordnungen die erreichte Fehlerquote, bei einer eingestellten Wahrscheinlichkeitsgrenze von P=0,98, mit etwa 20-30 % vergleichsweise relativ hoch. Innerhalb der entsprechend anderen Gruppe lag die Fehlerquote hingegen nur bei etwa 3-4%. Hauptursache für die erhöhte Fehlerquote ist vermutlich die Tatsache, dass die jeweiligen Zuordnungsmöglichkeiten der genannten fehleranfälligen Gruppe von NOESY-Signalen meist kurze (<0,5nm) und zugleich ähnliche Abstände innerhalb der Struktur aufweisen. Dadurch gleichen sich die Wahrscheinlichkeiten, für jede der Zuordnungsmöglichkeiten den Großteil des Signalvolumens zu erklären, entsprechend aneinander an. Dies führt allerdings zu einer Diskrepanz zwischen den Wahrscheinlichkeiten für diese spezielle Gruppe von Zuordnungsmöglichkeiten und den entsprechenden Wahrscheinlichkeiten innerhalb der Verteilungen. Das heißt, dass die vorhandenen Verteilungen für diese Zuordnungen weniger oder gar nicht mehr repräsentativ sind und es somit zu einer erhöhten Anzahl von falschen Zuordnungen kommen muss. Vor allem langreichweitige Zuordnungsmöglichkeiten haben innerhalb dieser Gruppe eine wesentlich höhere Wahrscheinlichkeit ein starkes NOESY-Signal zu erzeugen als aufgrund der Verteilungen bzw. statistisch zu erwarten ist. Dies erklärt, dass ein Großteil, meist über 50%, der falschen Zuordnungen an langreichweitige NOESY-Signale vergeben wird.

5.2.5.5 Unterschiedliche Anteile zugeordneter Signale bei verschiedenen Spektren

Wie bereits erwähnt, konnte die Anzahl der erstellten Zuordnungen für zwei- und dreideutige NOESY-Signale unter Einsatz der *neuen Datenbank 3* in etwa verdoppelt werden. Allerdings fiel auf (s. Kap 4.2.2.1), dass beim simulierten 2D-NOESY-Spektrum des Proteins HPr im Vergleich zum Spektrum vom Protein CSP meist etwa 10-14 % mehr der jeweils

vorhandenen zwei- und dreideutigen NOESY-Signale mit einer Wahrscheinlichkeit von mehr als 98% zugeordnet werden konnten. Dies war unabhängig vom jeweils eingestellten Suchradius. Das liegt höchstwahrscheinlich an den unterschiedlichen Größen der beiden Proteine mit jeweils 66 Resten für CSP bzw. je 88 Reste für das Protein HPr. Das Protein HPr entspricht aufgrund seiner Größe besser den innerhalb der Strukturdatenbasis vorkommenden Proteinen, welche eine durchschnittliche Sequenzlänge von 271 Aminosäuren besitzen. Das heißt, dass die interatomaren Abstände zwischen vergleichbaren Atomen im Vergleich zum CSP im Durchschnitt besser mit den entsprechenden Abständen der Proteine aus der Strukturdatenbasis übereinstimmen. Die vorhandenen NOESY-Signale im Spektrum vom Protein HPr wurden deshalb häufiger für eine der jeweils vorhandenen zwei- oder drei Zuordnungsmöglichkeiten als sehr Wahrscheinlich (hier 98 %) eingestuft. Generell ist somit zu erwarten, dass mit zunehmender Gleichheit des zu untersuchenden Proteins bezüglich der Durchschnittsgröße der Proteine innerhalb der Strukturdatenbasis, größere Anteile der jeweils vorhandenen zwei- und dreideutigen NOESY-Signale zugeordnet werden können. Dies zeigt allerdings auch, dass die Zuordnungsqualität bezüglich der Zuordnungsmenge unter Benutzung der neuen Verteilungen von den Eigenschaften des zu untersuchenden Proteins nicht vollständig unabhängig ist.

6.Ausblick

6.1 Erstellung individueller Datenbanken

Der Einsatz von Wahrscheinlichkeitsverteilungen bei dem hier beschriebenen Verfahren zur automatischen Zuordnung mehrdeutiger NOESY-Signale lässt sich noch weiter optimieren. Die im Rahmen dieser Arbeit generierten Datenbanken aus Wahrscheinlichkeitsverteilungen sind für die Zuordnung von NOESY Spektren unterschiedlicher Proteinen konzipiert worden. Optimaler wäre allerdings die Verwendung einer dem jeweils zu untersuchenden Protein individuell angepassten Datenbank. So könnte man beispielweise Wahrscheinlichkeitsverteilungen benutzen, in denen nur Atomabstände aus solchen Proteinstrukturen integriert wurden, welche mit dem jeweils zu untersuchenden Protein strukturell oder sequentiell verwandt sind. Hierbei ist zu erwarten, dass sich die Kurvenverläufe der resultierenden Verteilungen unterschiedlicher Atompaare im Vergleich zu den jetzigen Verteilungen in stärkerem Maße unterscheiden werden. Dies könnte auch für sequentiell weiter entfernte Atompaare gelten.

Der Grund für diese Annahme liegt darin, dass individuelle interatomare Abstände innerhalb des jeweils zu untersuchenden Proteins nicht, aufgrund vieler stark unterschiedlicher Proteinstrukturen, während der Berechnung von Wahrscheinlichkeitsverteilungen herausgemittelt werden. Bei Benutzung solcher Verteilungen ist somit zu erwarten, dass ein noch höherer Anteil von mehrdeutigen NOESY-Signalen eine Zuordnung mit jeweils einer hohen Wahrscheinlichkeit erhält. Dies könnte auch langreichweitige Signale betreffen. Allerdings muss hierbei für jedes neue zu untersuchende Protein eine eigene Datenbank aufgebaut werden. Dafür müsste die Erzeugung von Wahrscheinlichkeitsverteilungen aus bekannten Proteinstrukturen automatisiert werden bzw. für den Benutzer bequem durchzuführen sein. Dies erfordert zunächst die Zusammenführung der bislang noch separat vorliegenden für die Erstellung der Wahrscheinlichkeitsverteilungen benötigten Programme zu einem Programm. Dieses könnte dann in das Softwarepaket *AUREMOL* in Form einer neuen Funktion eingebaut werden.

6.2 Weitere Anwendungsmöglichkeiten der Datenbanken

Die in Rahmen der Arbeit erstellten Datenbanken aus Wahrscheinlichkeitsverteilungen enthalten große Mengen struktureller Information. Diese lässt sich, neben der automatischen Zuordnung von NOESY-Signalen, auch zur Optimierung anderer für die Proteinstrukturbestimmung wichtiger Arbeitschritte einsetzen. So ist es z.B. möglich auf der Basis von Abstandswahrscheinlichkeitsverteilungen die Gesamtenergie unterschiedlicher Proteinkonformationen oder Teile davon zu berechnen [79]. Dies wurde bereits für die Vorhersage lokaler Strukturen globulärer Proteine (M. Sippl 1989) erfolgreich durchgeführt. Aufgrund der nun vergleichsweise wesentlich größeren zu Verfügung stehenden Anzahl von Wahrscheinlichkeitsverteilungen, müsste das genannte Verfahren in entsprechend akkuraterer Form durchführbar sein. Kerngedanke dieses Verfahrens ist die Annahme, dass die Wahrscheinlichkeit für einen bestimmten Abstand zweier Atome in engen Zusammenhang mit dem Potential ihrer mittleren wechselwirkenden Kräften steht [79]. Das Energiepotential *E(s)* der Interaktion zweier Atome innerhalb eines Proteins lässt sich unter Anwendung des Gesetzes von Boltzmann als Funktion ihres Abstandes *s* berechnen über:

$$E(s) = -kT \ln[f(s)] - kT \ln[Z] \qquad (5.1)$$

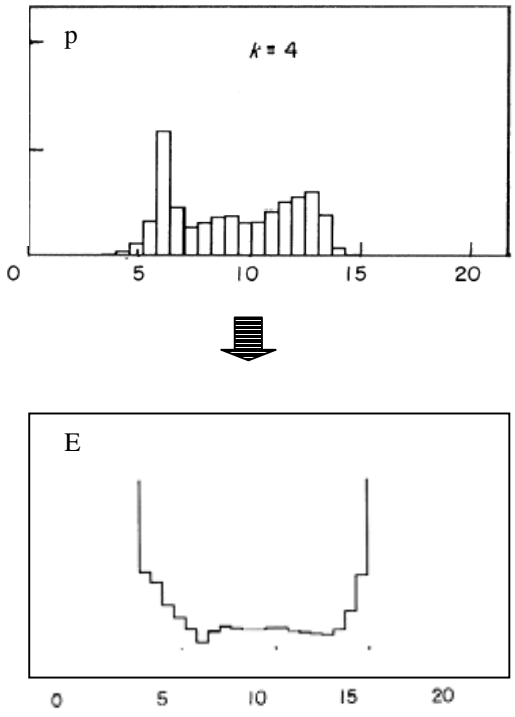


Abbildung 6.1: Transformation von abstandsabhängigen Wahrscheinlichkeitsdichten p zu abstandsabhängigen Energiepotentialen E. Hier schematisch gezeigt für Cα-Atompaare mit einem relativen Sequenzabstand (*k)* von je 4 Aminosäuren (nach M.J. Sippl [79])

Die Variable *f(s)* entspricht der Wahrscheinlichkeitsdichte für einen bestimmten Abstand *s* zweier Atome, *k* ist die Boltzmannkonstante und *T* die absolute Temperatur. Der Ausdruck $kT\ln[Z]$ ist hierbei eine Konstante wobei Z für die Boltzmannsumme [79] steht. In Abbildung 6.2 ist die Umwandlung von Abstandswahrscheinlichkeitsdichten zu entsprechenden Energiepotentialen veranschaulicht. Aus der Summe der Energiepotentiale mehrerer Atompaare kann nun die Gesamtenergie von bestimmten Bereichen oder der gesamten Konformation einer Proteinstruktur bestimmt werden. Damit ist es beispielsweise möglich die jeweils Energieärmsten aus den nach einer Strukturrechung resultierenden Konformationen eines Proteins zu ermitteln.

Weitere Anwendungen von Abstandswahrscheinlichkeitsverteilungen sind bei der Qualitätsbeurteilung sowie bei der Verfeinerung von Proteinstrukturen zu sehen. Es konnte bereits gezeigt werden, dass sich mit Hilfe solcher Verteilungen die native Struktur aus einer Menge von strukturell ähnlichen oder bereits energieminimierten Konformationen eines Proteins ermitteln lässt [80]. Hierbei macht man sich die Tatsache zu Nutze, dass in *„schlechten“* bzw. *„guten*“(=nativen) Strukturen verhältnismäßig viele bzw. wenige statistisch unwahrscheinliche interatomare Abstände vorkommen.

Abkürzungsverzeichnis

2D	Zweidimensional
3D	Dreidimensional
ANSI	American National Standards Institute
BSE	Bovine Spongiforme Enzephalopathie
CD	Circulardichroismus
CRINEPT	Cross Correlated Relaxation Enhanced Polarization Transfer
FID	Free Induction Decay
GB	Gigabyte
HSQC	Heteronuclear Single Quantum Coherence
IUPAC	International Union of Pure and Applied Chemistry
ISPA	Isolated Spin Pair Approximation
KB	Kilobyte
kDa	Kilo Dalton
MB	Megabyte
MD	Molecular Dynamics
MHz	Megahertz
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effekt
PDB	Protein Data Bank
ppm	Parts per million
TROSY	Transverse Relaxation Optimized Spectroscopy

Literaturverzeichnis

[1] Stryer Lubert (1996) *Biochemie* (4.Auflg.) Spektrum Akademischer Verlag.

[2] Withford David (2005) *Proteins Structure and Function* Wiley Verlag.

[3] Welsh Gary (2002) *Protein Biochemistry and Biotechnology* Wiley Verlag

[4] Lesk Arthur M. (2001*) Introduction to Protein Architecture* Oxford University Press.

[5] Voet D., Voet J.G. (1994) *Biochemie* VCH Verlagsgesellschaft.

[6] Horton H. Robert, Laurence A. Moran, Raymond S. Ochs, David J. Rawn, K. Gray Scr-imgeour *Principles of Biochemistry* (3.Auflage) Prentice Hall.

[7] Lehninger (2001) *Biochemie* (3.Auflage) Springer Verlag.

[8] Eriksson et al. *Nature*, **423** 293-298.

[9] Prusiner, S.B. (2001) *N.Engl.J.MED.,* **344**, 1516-1526.

[10] Ooms, F. (2000) *Current Medicinal Chemistry,* Vol 7, Number 2, 141-158(18).

[11] Baker D.and Sali A. (2001) *Science,* **5,** 93-96.

[12] Gutberlet *et al.* (2001) *Acta Cryst.,* **D57**, 349-354.

[13] Lottspeich F., Zorbas H., (1998) *Bioanalytik* Spektrum Akademischer Verlag.

[14] Byron, O., Gilbert, R.J.C. (2000) *Current Opinion in Biotechnology,* **11,** 72-80.

[15] Kelly, S.M., Price, N.C. (2000) *Current Protein and Peptide Science,* **1**, 349-384.

[16] Roman Tuma (2005) *Journal of Raman Spectroscopy,* 36(4):307.

[17] Alexey, S., Ladokhin (2000) *Encyclopedia of Analytical Chemistry,* pp. 5762-5779.

[18] van Heel M., Gowen B., Matadeen R., Orlova, E.V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M., Patwardhan, A., (2000*) Q Rev Biophys*., **33**, 307-69.

[19] Huber T & Torda A.E. (2002) (Tsiggelny, I.F.,ed,). *International University Line, La Jolla*, pp. 263-298.

[20] Bonneau R.& Baker D. (2001) *Annu. Rev. Biophys. Biomol Struct.,* **30**, 173-189.

[21] Marc, A., Renom Marti, Ashley, C., Fiser Andreas, Sanches Roberto, Melo Francisco and Sali Andrej (2000) *Annu. Rev. Biophys. Biomol. Struct.* **29**,291-325.

[22] Wüthrich, K. (2003) *ChemInform,* Volume 34, Issue 42 .

[23] Wüthrich, K. (2001) *Nature Structural Biology,* **8**, 923 – 925.

[24] Bax, A., Grzesiek, S*., Acc. Chem. Res*., 1993, **26**, 131-138.

[25] Riek, R., Pervushin K., & Wüthrich, K. (2000) *Trends Biochem. Sci.,* **25**, 462-468.

[26] Pervushin, K. et al. (1997) *Proc. Natl. Acad. Sci.,* U. S. A. **94**, pp. 12366–12371.

[27] Riek, R. et al. (1999) *Proc. Natl. Acad. Sci.,* U. S. A. **96** pp. 4918–4923.

[28] Ernst, Richard, R., Wokaun., Alexander, Bodenhausen, Geoffrey (1990) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions***,** Oxfort University Press.

[29] Hausser, Karl H., Kalbitzer, Hans R. (1989) *NMR für Mediziner und Biologen. Strukturbestimmung, Bildgebung, In-vivo- Spektroskopie*, Springer-Verlag GmbH.

[30] Gronwald, W., Brunner, K., Kirchhöfer, R., Nasser, A., Trenner, J., Ganslmeier, B., Riepl, H., Ried, A., Scheiber, J., Elsner, R., Neidig, K.P., Kalbitzer, H.R. (2004) *Bruker Reports*; **154/155** 11-14.

[31] XWINNMR Program, *Bruker, Biospin GmbH*, Ettlingen (2003).

[32] Herrmann, T., Güntert, P. and Wüthrich, K. (2002) *J. Biomol. NMR,* **24** p. 171.

[33] Nilges, M. (1995) *J. Mol. Biol.,* **245** p. 645.

[34] Duggan, B.M., Legge, G.B, Dyson, H.J., and Wright P.E. (2001) *J. Biomol. NMR,* **19** p. 321

[35] Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein R. and Thornton J. M. (1996) *J. Biomol. NMR,* **8** p. 477.

[36] Kraulis, P.J. (1989) *J. Magn. Reson.,* **84** p.627.

[37] Neidig, K.P., Geyer, M., Görler, A., Antz, C., Saffrich, R., Beneicke, W., and Kalbitzer H.R. (1995) *J. Biomol. NMR* **6** p.255.

[38] FELIX Program, *Accelrys Inc*., San Diego, CA (2003).

[39] Kremer, W., Schuler, B., Harrieder, S., Geyer, M., Gronwald, W., Welker, C., Jaenicke, R., u. Kalbitzer, H. R. (2001) *Eur. J. Biochm.,* **268**,2527-2539.

[40] Görler, A., Hengstberg, W., Kravanjia, M., Beneicke, W., Kalbitzer, H.R. (1999) *Appl. Magn. Reson.*, **17**, 465-480.

[41] Rost, B., Schneider, R. and Sander, C., J.(1997) *J. Mol. Biol*., **270**, 471-480.

[42] Ried, A., Gronwald, W., Trenner, J.M., Brunner, K., Neidig, K.P. & Kalbitzer, H.R. (2004) *J. Biomol. NMR*, **30** 121-131.

[43] Word et al.(1999).*J. Mol.Biol.,* **285,** 1733-1747.

[44] Bronstein, I.N., Semendjaew, K.A., Musiol, G., Mühlig, H. (1995) *Taschenbuch der Mathematik*, Verlag Harri Deutsch.

[45] Evans, Jeremy N.S.(1995) *Biomolecular NMR Spectroscopy*, Oxford University Press.

[46] H. Press William, Flannery Brian P., Teukolsky Saul A., Vetterling William T.(1992) *Numerical Recipes in C*, Cambridge University Press.

[47] http://www.rcsb.org.

[48] Koradi, R, Billeter, M., Engeli, M., Guntert, P. and Wüthrich, K. *J.* (1998) *Magn. Reson.* **135**, p. 288.

[49] Gronwald, W., Willard, L., Jellard, T., Boyko, R.F., Rajarathnam, K., Wishart, D.S., Sönnichsen F.D., and Sykes, B.D. (1998) *J. Biomol. NMR* **12**, p. 395.

[50] Bartels, C., Billeter, M., Güntert, P., and Wüthrich, K. (1996) *J. Biomol. NMR* **7**, p. 207.

[51] Hitchens, T.K., Lukin, J.A., Zhan, Y., McCallum, S.A. and Rule, G.S. (2003) *J. Biomol. NMR* **25** , p.1.

[52] Herrmann,T., Güntert, P. and Wüthrich, K. *J. Mol. Biol.* **319** (2002), p. 209.

[53] Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Trenner, J.,Kremer, W., Neidig, K.P. and Kalbitzer, H.R. *J. Biomol. NMR* **23** (2002), p. 271.

[54] Case, D.A., Pearlman D.A., Caldwell, J.W., Cheatham, T.E., III, Wang, J., Ross, W.S., Simmerling, C.L., Darden, T.A., Merz, K.M., Stanton, R.V., Cheng, A.L., Vincent, J.J., Crowley, M., Tsui V., Gohlke, H., Radmer, R.J., Duan, Y., Pitera, J., Massova, I., Seibel, G.L. Singh, U.C., Weiner, P.K. and Kollman, P.A. (2002).*AMBER 7 Program*,University of California, San Francisco, CA.

[55] Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grossekunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S.,Read, R.J., Rice, L.M., Simonson,T. and Warren, G.L. (1998) *Acta Cryst.* **D54**, p. 905.

[56] Herrmann, T., Güntert, P. and Wuthrich, K. (2002) *J. Mol. Biol.* **319**, p. 209.

[57] Gronwald, W., Kirchhöfer, R., Görler, A. Kremer, W., Ganslmeier, B., Neidig, K.P. and Kalbitzer, H.R. (2000) *J. Biomol. NMR* **17**, p. 137.

[58] Sippl, M.J. (1993) *Proteins* **17**, p.355.

[59] Wayne Boucher (2002), AZARA Program, Department of Biochemistry, University of Cambridge.

[60] TRIAD Program (2003), Tripos Inc., St Louis, MO.

[61] Shaw, M.K. & Ingraham, J.L. (1967) *J. Bacteriol.* **94**, 157–164.

[62] Schindelin, M. & Heinemann, U. (1994) *Proc. Natl Acad. Sci.* USA **91**, 5119–5123.

[63] Newkirk, K., Feng, W., Jiang, W., Tejero, R., Emerson, S.D., Inouye, M. & Montelione, G.T. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 5114–5118.

[64] Schindelin, H., Mahariel, M.A. & Heinemann, U. (1993) *Nature (London)* **364**, 164–168.

[65] Schnuchel, A., Wiltschek, R., Czisch, M., Herrler, M., Willimsky, G., Graumann, P., Mahariel, M.A. & Holak, T.A. (1993) *Nature (London)* **364**, 169–171.

[66] Mueller, U., Perl, D., Schmid, F.X. & Heinemann, U. (2000) *J. Mol. Biol.* **297**, 975–988.

[67] Murzin, A.G. (1993) *EMBO J.* **12**, 861–867.

[68] Maurer, T., Meier, S., Kachel, N., Munte, C.E., Hasenbein, S., Koch B., Hengstenberg, W., Kalbitzer, H.R. (2004) *J. Bacteriol..,* **186** (17) 5906-18.

[69] Anderson, B., Weigel, N., Kundig, W., Roseman S. (1971) *J. Biol. Chem.* **246**,7023-7033.

[70] Hengstenberg, W., Penberthy, W.K., Hill, K.L., Morse, M.L. (1969*) J. Bacteriol*. **99**, 383-388.

[71] Jaffor, Ullah , A.H., Cirillo, V.P. (1976) *J.Bacteriol.* **127**, 1298-1306.

[72] Kalbitzer, H.R., Hengstberg, W., Rösch, P., Muss, P., Bernsmann, P., Engelman, R., Dörschung, M., Deutscher, J. (1982) *Biochemistry* **21**,2879-2885.

[73] Marquet M., Creignou, M.C., Dedonder, R. (1976) *Biochemistry* **58**, 435-441 (1976).

[74] Beneski, D.A., Nakazawa, A., Weigel, N., Hartman, P.E., Roseman, S. (1982) *J.Biol.Chem.* **257**,14492-24498.

[75] Deutscher, J., Pevec, B., Beyreuther, K., Kiltz, H.-H., Hengstenberg, W. (1986) *Biochemistry* **25**, 6543-6551.

[76] Reizer J., Peterkofsky, A., Romano, A. (1988) *Proc. Natl. Acad.Sci*. USA 85,2041-2045.

[77] Kruse, R., Hengstenberg, W., Beneicke, W., Kalbitzer, H.R. (1993) *Protein Eng*. **6**,417-423.

[78] Postma PW. et al, (1993) *Microbiol. Rev.* **57** 543-594.

[79] Sippl, M., (1990) *J. Mol. Biol.*, **213** 859-883.

[80] Subramaniam Shankar, Tcheng David, K., Fenton, James M. (1996) *ISMB*: 218-229.

[81] Ganslmeier, B. (2002), *Sofwareprojekt zur automatischen Auswertung von multi-dimensionalen NMR-Spektren*, Dissertation, Regensburg.

[82] Möglich Andreas, Weinfurtner Daniel, Gronwald Wolfram, Maurer Till, Kalbitzer Hans Robert (2005),*Bioinformatics,* Volume 21, **9**, 2110 – 2111.

[83] Trenner, J. M. (2006), *Accurate proton-proton distance calculation and error estimation from NMR data for automated protein structure determination in AUREMOL,* Dissertation, Regensburg.

[84] Boelens, R., Koning, M.G., van der Marel, G.A., van Boom, J.H., Kaptein, R., (1989) *J. Magn.Reson.* **82** 290.

[85] Borgias, B.A, James, T.L.,(1990) *J.Magn.Reson.***87** 475.

[86] Post,C.B,., Meadows, R.P., Gorenstein, D.G.,*J.Am.Chem.Soc.*112.

[87] van de Ven, F.J.M., Blommers, M.J.J., Schouten, R.E., Hilbers, C.W., (1991) *J.Magn.Reson.***94** 140.

[88] Madrid, M., Llinas, E. (1991) *J. Magn.Reson.***93** 329.

[89] Kim, S.G., Reid, B.R (1992) *J.Magn.Reson.***100** 383.

[90] Güntert, P., Mumnethaler, C., Wüthrich, K. (1997 Oct 17) *J.Mol.Biol.*; **273**(1) 283-98.

[91] Güntert, P., Braun, W., Wüthrich, K. (1991) *J.Mol.Biol.* **217** 517.

[92] Mumenthaler C., Braun W. (1995) *J.Mol.Biol* **254** 465.

[93] Güntert P. (2003) *Prog. NMR Spectrosc.* **43**, 105-125.

[94] B.A., Johnson, R.A., Blevins (1994) *J.Biomol NMR* 4603.

[95] Bartels, C., Xia, T., Billeter, M., Güntert, P., Wüthrich, K. (1995), *J.Biomol NMR* **6** 1.

[96] Huang, Y. J., Tejero, R., Powers, R., Montelione, G.T. (2006) *Struct. Funct. Bioinformatics* **15**, 587-603.

[97] Huang, Y. J., Powers, R., Montelione, G.T. (2005),*Chem Soc*. **127**,1665-1674.

[98] Grishaev, A., Llinas, M. (2002) *Proc. Natl. Acad. Sci.* USA **99.**

[99] Tejero, R., Monleon, D., Celda, B., Powers, R. and Montelione, G.T. (1999) *J. Biomol. NMR* **15** 251-26.

[100] Görler, A., Antz, C., Neidig, K. P., Kalbitzer, H. R. (1997) *J. Magn. Reson.* **129** (2),165-172 .

[101] http://www.winzip.de.

Anhang

A. Liste aller Wasserstoffatomnamen in den 20 natürlichen Aminosäuren (nach IUPAC)

Aminosäure	Wasserstoffatome
Alanin	HN,HA,HB1,HB2,HB3
Arginin	HN,HA,HB2,HB3,HG2,HG3,HD2,HD3,HE,HH11,HH12 ,HH21,HH22
Asparagin	HN,HA,HB2,HB3,HD21,HD22
Aspartat	HN,HA,HB2,HB3,HG
Cystein	HN,HA,HB2,HB3,HG
Glutamin	HN,HA,HB2,HB3,HG2,HG3,HE21,HE22
Glutamat	HN,HA,HB2,HB3,HG2,HG3,HE2
Glycin	HN,HA2,HA3
Histidin	HN,HA,HB2,HB3,HD2,HE1,HD1,HE2
Isoleucin	HN,HA,HB,HG12,HG13,HG21,HG22,HG23,HD11,HD1 2,HD13
Lysin	HN,HA,HB2,HB3,HG2,HG3,HD2,HD3,HE2,HE3,HZ1, HZ2,HZ3
Methionin	HN,HA,HB2,HB3,HG2,HG3,HE1,HE2,HE3
Phenylalanin	HN,HA,HB2,HB3,HD1,HD2,HE1,HE2,HZ
Prolin	HA,HB2,HB3,HG2,HG3,HD2,HD3
Serin	HN,HA,HB2,HB3,HG
Threonin	HN,HA,HB,HG1,HG21,HG22,HG23
Tryptophan	HN,HA,HB2,HB3,HD1,HE1,HE3,HZ2,HZ3,HH2
Tyrosin	HN,HA,HB2,HB3,HD1,HD2,HE1,HE2,HH
Valin	HN, HA, HB,HG11,HG12,HG13,HG21,HG22,HG23

B. Benutzte Strukturdatenbasis (PDB-Datei-Codes)

1	1aa0	1xnb	1dek	1znj	1rbo	1ako	2dri	1aqz	2ifm	2mev	1msk	1nbc
2	1ctf	1abo	1iso	1dok	2ace	1edg	1aos	1fc2	1atl	1awd	1tab	1ten
3	1iae	1daa	1ptf	1jhg	1ais	1kuh	1etb	1lt3	1fit	1fok	2ovo	2psp
4	1aa2	1ihf	1ytf	1agj	1eag	1rgs	1lit	1sfe	1maz	1mjc	1bba	1bdo
5	1ctn	1ppt	1aep	1dor	1klo	1akz	1ryt	2gdm	1smn	1sri	1msp	1gbs
6	1pne	1xsm	1der	1jkw	1rcb	1edh	2drp	1ar5	2ifo	2mhr	1taf	1tf4
7	1wod	1abr	1isu	1pyi	2ach	1kum	1etp	1fca	1fjl	1awj	2pde	1beb
8	1aa6	1dai	1pth	1agq	1ebd	1rhs	1lki	1lt5	1mbd	1fos	1bbh	1gca
9	1cto	1ihv	1ytw	1dos	1knb	2bgu	1sac	1sft	1smp	1mka	1fwp	1ncx
10	1wtu	1prc	1aer	1jli	1rcf	1al0	2ebn	2gf1	2ilk	2min	1mty	1tfe
11	1aa7	1xso	1dhp	1pys	2acy	1edm	1aoz	1arb	1au7	1ax4	1tag	1bec
12	1ctt	1abv	1itb	1agr	1aj8	1kve	1eut	1fcd	1fkj	1fow	2pgd	1gcb
13	1idk	1dap	1ptq	1dpe	1ebp	1rie	1lkk	1lta	1mbg	1mla	1bbl	1ndh
14	1poa	1il6	1af5	1jly	1kny	2blt	1sap	1sh1	1smr	1srs	1fxd	2qil
15	1x11	1pre	1dhr	1pyt	1rcy	1alk	2end	2gli	2kau	2mnr	1muc	1beo
16	1cvl	1xva	1itg	1ah7	2adm	1efn	1exg	1lts	1aui	1frb	1taq	1gcm
17	1ido	1dar	1pty	1dpg	1ajj	1kzu	1lkt	1shc	1fkx	1mml	2phl	1neq
18	1poc	1imb	1yui	1jmc	1ece	2bnh	1sat	2gmf	1mdy	1std	1fyc	1tfr
19	1xaa	1prn	1af7	1qap	1kob	1alo	2eng	1arv	1smt	2mpr	1mup	2reb
20	1aac	1xxa	1dhx	1drw	1rec	1eft	1exn	1fdl	2lbd	1axi	1tbg	1bfg
21	1cwp	1ac5	1ith	1jpc	2af8	1l92	1lla	1luc	1fle	1fre	2phy	1gdf
22	1idy	1dbq	1puc	1qba	1ajq	1ris	1sbp	1sig	1mfa	1stf	1bbp	1neu
23	1xbr	1xyz	1yve	1aho	1ecf	2bop	2erl	2hbg	1smv	2mrb	1fzb	1tgx
24	1cxs	1dbr	1dik	1dup	1kpt	1aly	1ext	1ash	1av1	1fro	1mxb	2rn2
25	1iea	1inp	1ivy	1jsu	1reg	1efu	1lmb	1lvk	1flp	1mof	1tbr	1gdh
26	1pot	1prt	1pud	1qbe	2aop	1lam	2ezh	2hft	1mh1	1stm	2pia	1the
27	1xel	1yas	1din	1ahs	1ajs	1rlr	1apy	1ass	1snc	2mta	1bbt	2rsl
28	1aay	1dcp	1ixh	1dut	1krn	2bpa	1fap	1fdr	2liv	1frp	1g3p	1bgc
29	1cyd	1iph	1pue	1qli	1req	1amj	1lml	1lxa	1avc	1mol	1myl	1gdo
30	1ifi	1pru	1afr	1zxq	2arc	1efv	1scu	1skz	1fmb	2myr	1tc3	1nfk
31	1pov	1ycc	1div	1aie	1ecl	1lay	2ezk	2hnt	1mhl	1ayl	2pii	1thg
32	1xer	1ad2	1iyu	1dxg	1krs	1rmd	1aq6	1asz	1sp2	1frv	1bcf	2rsp
33	1cyj	1ddf	1afv	1qnf	2asi	2brd	1far	1fds	2ltn	1mpg	1mzm	1gen
34	1igd	1ir3	1djx	1zym	1ak0	1amk	1lpb	1lyl	1avo	1svb	1tca	1nfn
35	1pox	1psc	1jac	1aih	1ecm	1lba	1sdf	1slt	1fmk	2nad	2plc	1thj
36	1xgs	1ycq	1pvc	1dxy	1rfb	1rmv	2fal	2hpd	1mhu	1aym	1bco	2sas
37	1ab8	1ad3	1afw	1jxp	2ayh	2btf	1aqb	1at0	1spb	1mrj	1gai	1bgl
38	1cyo	1ddt	1dkg	1qor	1ak1	1egp	1fba	1ffh	2mas	1svp	1nal	1ghf
39	1ign	1irf	1jdc	256b	1ecp	1lbd	1se4	1lyp	1avp	2nll	2pol	1nfp
40	1ppb	1psd	1pvd	1aik	1rfs	1rnl	2fha	1slu	1fmt	1ayn	1bcp	1thm
41	1xik	1ycs	1zin	1dyn	2bb2	2cas	1aqe	2hts	1mhy	1fsu	1gal	2scp
42	1ab9	1ade	1ag8	1kay	1ak5	1amm	1fbr	1atb	2mbr	1mrp	1nar	1bgp
43	1cyx	1dea	1dkz	1qrd	1ecr	1ehs	1lrv	1fie	1avy	1svr	1tcr	1ghj
44	1igs	1iro	1jdw	2aaa	1kte	1lbe	1sei	1lzr	1fnc	2ohx	2por	1tht
45	1ppn	1yge	1pvu	1ail	1rge	1rom	2fua	1sly	1mil	1bab	1gar	2sga
46	1xjo	1adj	1znb	1kit	2bbk	2cba	1aqt	2i1b	1sqc	1fur	1nba	1bgw
47	1aba	1irs	1agd	1r69	1ak6	1amp	1fc1	1ati	2mcm	1msc	1tdt	1ghr
48	1d66	1yrn	1dmc	2abk	1ede	1lbu	1lst	1fip	1fnf	2omf	2prd	1nhp
49	1igt	1ytb	1jev	1air	1rgp	1an2	1ses	1mai	1mit	1bam	1bdb	1thv
50	1ppr	1ae9	1pya	1eaf	2bbv	1ema	2fxb	1smd	1sra	1fvk	1gat	2sil

51	1bhg	1gpc	1tup	1oxa	1vie	1clc	1hus	1pyc	1axh	1pft
52	1gif	1tnr	3grs	1hcr	4mt2	1hrd	1wgj	1zto	1mnt	1vpu
53	1nif	1bnd	1btn	1oyc	1cfr	1phd	9rnt	1zwa	1axj	1hqi
54	1thx	1gpl	1guq	3tgl	1hlo	1vsg	1cpt	1zwb	1ayj	1vtx
55	2spc	1nul	1onr	1cdy	1pdo	5rub	1huw	1zwc	1stu	1pls
56	1bi6	1bor	3ink	1hcz	1vih	1cmb	1pma	1zwd	1fsb	1whf
57	1gky	1gpm	1opc	1pam	4pga	1hry	9wga	1ah9	1sxl	1pmc
58	1nip	1nwp	3lad	3ull	1chc	1php	1crk	1zwf	1fvl	
59	1tif	1toc	1gvp	1cei	1pdr	6fd1	1hxn	1zwg	1bbo	
60	1bia	2yhx	1opr	1pau	1vii	1cns	1whi	1jun	1gab	
61	1glc	1bov	1tvx	1v39	4pgm	1hst	1csh	1jvr	1tcp	
62	1tig	1gpr	1gw4	451c	1chd	1phr	1hyh	1e2b	1bct	
63	2stv	1nzy	1orc	1cel	1hmt	1vvc	1pmi	1kjs	1ncs	
64	1bif	1tpg	1tys	1pax	1pea	6ins	1who	1aj3	2pta	
65	1gln	3bct	3mde	1vca	1vin	1cnt	1csn	1eci	2ptl	
66	1nls	1bp1	1gyp	4aah	4rhn	1htm	1hyp	1ajy	1tfi	
67	1tii	1oac	1ord	1cem	1chk	1pi2	1wht	1ret	1bfm	
68	1gnd	1trb	1tyv	1hfc	1hmy	1wab	1hyt	1ksr	1nfa	
69	1nnc	3chy	3min	1pbe	1ped	7ahl	1pmy	1kst	1bgk	
70	1tit	1bpi	1gzi	1vcc	1vjs	1cnv	1wit	1ktx	1ngr	
71	2tbv	1grj	1ort	4cpa	4rhv	1htn	1pnb	2bds	2stt	
72	1ble	1obp	1cax	1ceo	1chm	1pii	1wkt	1rip	1nkl	
73	1gnw	1trk	1ha1	1hfh	1hnf	1waj	1iba	1egf	2tbd	
74	1nox	3cla	1osa	1pbg	1vls	7api	1ica	1aml	1bip	
75	2tct	1bpy	3nul	1vdf	4sgb	1cof	1pnh	1eit	1tiv	
76	1bmd	1occ	1cb2	4enl	1cid	1htp	1aab	1ron	1bnb	
77	1gof	1tsg	1osp	1cew	1hnr	1pjr	1pon	2cbh	1ntr	
78	1npk	3cox	3pch	1hge	1pfi	1wap	1aaf	1roo	2u1a	
79	1tlf	1brh	1cbn	1pbn	1vmo	7fab	1ab3	1ldr	1tnt	
80	2tgi	1gsa	1han	1vhh	4xis	1col	1abz	1erd	2vil	
81	1bmf	1ofg	1otf	4gat	1cii	1htr	1iml	2crd	1gpt	
82	1gow	1tuc	3pfk	1cex	1hoe	1pkm	1prs	1rtn	1grx	
83	1npo	3cyr	1cby	1hgx	1pfs	1wba	1aca	1rtn	1tum	
84	1tlk	1bro	1hav	1pbw	1vnc	7rsa	1acp	1ery	1tvs	
85	2tmd	1gtp	1otg	1vhi	5csm	1coo	1pse	1aoy	1gur	
86	1bmt	1ois	3pmg	4gcr	1ciq	1huc	1dec	1ap8	1tvt	
87	1gox	1tud	1cc5	1cfb	1vol	1pkp	1psm	1apj	3leu	
88	1nqb	3dni	1hce	1hiw	5eat	1wdc	1adn	1apq	3mra	
89	1tmf	1brs	1oun	1pda	1ciy	8abp	1def	1sco	1hae	
90	2trc	1gtq	3pte	1vhr	1hpi	1cpc	1isk	1lre	3tgf	
91	1bmv	1olg	1cd8	4htc	1pfx	1hul	1psv	1ark	1cds	
92	1gp1	1tul	1hcg	1hjr	1vps	1plc	1yua	1fct	1hdj	
93	1nsj	3ebx	1ova	1pdc	5icb	1wer	1afo	1fdm	1hev	
94	1tml	1btk	3sdh	4ifm	1cka	8atc	1zaq	1sis	1cfe	
95	2ts1	1gtr	1cdc	1hlb	1hpt	1cpo	1afp	1aty	1vib	
96	1gpb	1omp	1hcl	1pdg	1pgs	1hur	1zdc	1auu	1cfh	
97	1nsy	3fib	1ovb	1vid	1ckm	1plq	1put	2leu	1cfp	
98	1tmy	1btl	3ssi	4kbp	1pgt	1wfb	1zec	1spf	1hma	
99	2tys	1gua	1cdk	1hle	1vsd	8fab	1zfd	1aw0	1peh	
100	1bnc	1onc	1hcn	1pdn	5ptp	1cpq	1agg	1sro	1hph	

Röntgen-Strukturen **NMR-Strukturen**

In der Tabelle sind die PDB-Codes aller bei der Erzeugung der Datenbanken benutzten Proteinstrukturen aufgeführt**.**

C. Charakteristische interatomare Atomabstände innerhalb von Sekundärstrukturen

Abstand	α̃Helix	310-Helix	β–Faltblatt (antiparallel)	β–Faltblatt (parallel)	Typ-I-Schleife	Typ-II-Schleife
d αn(i,i)	2.7Å	2.7Å	2.8Å	2.8Å	2.8Å	2.7Å/2.2Å
d αn(i,i+1)	3.5Å	3.4Å	2.2Å	2.2Å	3.4Å/3.2Å	2.2Å/3.2Å
d nn(i,i+1)	2.8Å	2.6Å	4.3Å	4.2Å	2.6Å/2.4Å	4.5Å/2.4Å
d nn(i,i+2)	4.2Å	4.1Å			3.8Å	4.3Å
d αn(i,i+3)	3.4Å	3.3Å			3.1Å-4.2Å	3.8Å-4.7Å
d αn(i,i+4)	4.2Å					

Die Tabelle zeigt typische Abstände *d* (Å) wie sie in Sekundärstrukturen innerhalb von Proteinen zwischen den Wasserstoffatomen *HA* (*α*) und *HN* (*n*) gefunden werden [45]. Der Buchstabe *i* kennzeichnet jeweils die Sequenzposition des Restes in dem sich das erste der beiden Atome befindet.

Danksagung

Sehr herzlich möchte ich mich bei Prof. Dr. Dr. Hans Robert Kalbitzer bedanken, der mir die Möglichkeit gab, in einem so fächerübergreifenden Gebiet, wie der Biophysik, zu promovieren. Trotz der oft knapp bemessenen Zeit als Institutsleiter, hat er sich stets sehr viel Zeit für ausführliche Diskussion genommen, die oft entscheidende Anstöße für den Fortschritt und Gelingen meiner Arbeit gaben.

Besonders möchte ich mich bei meinem Betreuer PD Dr. Wolfram Gronwald für seine große Hilfsbereitschaft und der fachkompetenten Betreuung meiner Arbeit bedanken.

Bedanken möchte ich mich bei Dipl. Math. Frank Braun für seine Beratung bei der Wahl geeigneter mathematischer Methoden.

Für die Überprüfung der Arbeit auf Rechtschreib- bzw. Kommafehler möchte ich mich herzlich bei meiner ehemaligen Stockwerksgenossin Charlotte Ahlswede bedanken.

Für die angenehme Arbeitsatmosphäre bedanke ich mich bei meinen Zimmerkollegen Dr. Jochen Trenner und Dr. Thorsten Graf sehr herzlich.

Besonderer Dank gilt meinen Eltern, die es mir durch ihre finanzielle Unterstützung ermöglicht haben, zu studieren und zu promovieren.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig angefertigt habe und keine, außer den angegebenen Hilfsmitteln, verwendet habe.

Regensburg, im Dezember 2006

(Adel Nasser)